# Asset Embeddings

Xavier Gabaix Ralph S.J. Koijen Robert J. Richmond Motohiro Yogo[*]

September 14, 2023. Preliminary and incomplete

**Abstract**

Firm characteristics are ubiquitously used in economics. These characteristics are often based on readily-available information such as accounting data, but those reflect only a part of investors' information set. We show that useful information about firm characteristics is embedded in investors' holdings data and, via market clearing, in prices, returns, and trading data. Based on insights from the recent artificial intelligence (AI) and machine learning (ML) literature, in which unstructured data (e.g., words or speech) are represented as continuous vectors in a potentially high-dimensional space, we propose to learn asset embeddings from investors' holdings data. Indeed, just as documents arrange words that can be used to uncover word structures via embeddings, investors organize assets in portfolios that can be used to uncover firm characteristics that investors deem important via asset embeddings. This broad theme provides a natural bridge to connect recent advances in the fields of AI and ML to finance and economics. Specifically, we show how *language* models, including transformer models that feature prominently in large language models such as BERT and GPT, can handle *numerical* information, and in particular holdings data to estimate asset embeddings. We provide initial evidence on the value added of asset embeddings through a series of applications in the context of firm valuations, return comovement, and uncovering asset substitution patterns. As a by-product, the models generate investor embeddings, which can be used to measure investor similarity. We propose a programmatic list of potential applications of asset and investor embeddings to finance and economics more generally.

[*] xgabaix@fas.harvard.edu, ralph.koijen@chicagobooth.edu, rrichmon@stern.nyu.edu, myogo@princeton.edu. For comments and suggestions, we thank John Campbell, Tengyuan Liang, Tarun Ramadorai, and seminar participants at the Chicago Booth machine learning working group, the NY Fed, the London Quant Group 2023 Autumn Seminar, and the NBER SI 2023 Forecasting & Empirical Methods conference. We thank Manav Chaudhary for excellent research assistance. For financial support, Gabaix thanks the Ferrante Fund, and Koijen thanks the Center for Research in Security Prices at the University of Chicago and the Fama Research Fund at the University of Chicago Booth School of Business.

# 1 Introduction

The success of embedding models to analyze text, image, and audio, has revolutionized the understanding of complex unstructured data by representing data (e.g., words or speech) as continuous vectors in a potentially high-dimensional space. Despite the success of embedding techniques in these fields, their application to the realm of finance and economics remains largely unexplored. Our main contribution is to introduce "asset embeddings" and show, both empirically and theoretically, that portfolio holdings (and transformations thereof) are particularly well suited to estimate asset embeddings. We illustrate how those asset embeddings allow economists to go beyond the common practice of using observable characteristics based on, for instance, accounting data.

A text embedding is a numerical representation of text data that captures the meaning of words or phrases in a vector space, and is usually extracted from data such as text via natural language processing. Likewise, we define an asset embedding as a numerical representation of the characteristics of an asset or firm in a vector space, that is extracted from data such as holdings and trading data (as we do in this paper), or other sources of non-traditional data (e.g., text). This contrasts with the more conventional measure of firm characteristics based on readily-available data, such as accounting data.

To estimate asset embeddings, we rely on embedding data sets. Embedding data sets have two dimensions, for instance, daily returns across firms within a quarter or text data across 10-K filings for firms. Our framework is flexible and can accommodate data from a variety of sources, such as prices, returns, volume or even text data. Given this flexibility, we develop a simple equilibrium asset pricing model in Section 3 to argue that holdings data (and transformations thereof) are particularly well suited to estimate asset embeddings. Intuitively, investors choose portfolios based on characteristics that capture a firm's risk and expected returns, and non-pecuniary benefits and costs such as a firm's sustainability. Put simply, documents structure words in ways that allow us to estimate word embeddings and, analogously, portfolio managers organize securities in ways that allow us to estimate asset embeddings. This simple insight enables us to connect to many of the recent modeling advance in machine learning and artificial intelligence, with asset embeddings being an important first step.

We introduce a general framework to estimate asset embeddings in Section 2. As the data that we use is typically significantly smaller than the data used in the context of text, image or audio models, we consider two types of asset embeddings.[1] Unsupervised embeddings are general-purpose embeddings that are estimated without taking into account how the embeddings are used

[1] As we discuss in Section 3.2, there are several advantages of our setting as the quality of the data are higher than text data and there are fewer stocks than words. Also, the latest generation of language models estimate very high-dimensional embeddings that are well beyond the number of observable characteristics used in finance and economics. When scaled by either the number of stocks or the number of parameters, the holdings data is of a similar order of magnitude as the text data used to estimate these models.

in applications. Supervised embeddings, by contrast, take into account how the embeddings are used. By taking into account how embeddings are used, smaller data sets can be used to estimate asset embeddings.[2]

We implement that insight for a variety of machine learning models, starting with a recommender system, to more advanced models such as Word2Vec, and the recent transformer architectures that are used in large language models such as BERT and GPT. An important question is how this modeling architecture can be used with *numerical* values instead of discrete *words*. That is, how to transform numerical information on holdings into BERT-readable language, which uses words. We propose such a bridge, and we estimate a model following the BERT architecture that we label "AssetBERT". In general, we estimate the model based on ranked portfolio holdings; indeed, we replace the ranking of words in a sentence by the ranking of stocks in a portfolio. In case of BERT, we use a masked language task: we predict the identity of e.g. the 10th largest holdings of a given investor, given data on the other holdings (see Figure 1 below). This is analogous to the completion tasks on which BERT was trained, where BERT has to guess the missing word, in a sentence like "The Fed decided to _____ rates to fight inflation". This broad theme, which can be applied to other modeling approaches in the audio, vision, and text literature, provides a natural bridge to connect recent advances in the fields of AI and ML to finance and economics.

To make the analogy more precise, it is useful to think of portfolio holdings as providing three types of information: (i) which, say, 60 stocks does an investor hold out of thousands of stocks, (ii) how does the investor rank these 60 stocks in its portfolio (e.g., Apple is the largest position), and (iii) how much does the investor allocate to these ranked positions (e.g., the largest position is 2% larger than the second-largest position). In our mapping to language models, we use the extensive margin (which stocks does the investor hold) and the relative positioning of the 60 stocks. In our current implementation of nonlinear language models, we do not use the information about the magnitudes of the allocations, only the ranks. In future versions, we can explore such extensions, but by using ranks, we can use off-the-shelf software that is optimized to handle massive amounts of data and estimate models on state-of-the-art hardware, in particular GPUs and TPUs. To assess whether the loss of information is large, we explore simpler PCA-style models using (log) holdings and ranked holdings, and we find them to perform similarly in our applications. Of course, we can reduce the information used to only the extensive margin (a "bag of words" approach), but then we unnecessarily remove information about the relative positions in an investor's portfolio. Our approach therefore uses as much information as possible, while being able to take advantage of the large and rapidly-developing NLP literature and accompanying software ecosystem.

We provide several preliminary implementations using data for the period from 2000.Q1 to 2021.Q4. We consider three applications focused on explaining firm valuations, the co-movement

[2] This logic is also applied in the context of empirical asset pricing by Lettau and Pelger (2020) and Bryzgalova et al. (2023).

of stock returns, and asset substitution patterns. When explaining valuations in the cross-section (both in- and out-of-sample) with either observable characteristics (we use those included in the demand system of Koijen and Yogo (2019)) or asset embeddings from various recommender systems and Word2Vec models, we find that asset embeddings perform on par with observable characteristics, and oftentimes outperform them. This conclusion holds period-by-period and the model's ability to explain valuations is stronger in the second part of our sample. We use fairly aggregated holdings data from regulatory 13F filings for our main results, and our results strengthen with more disaggregated data from mutual funds, exchange-traded funds (ETFs), closed-end funds, and variable annuity funds. In ongoing work, we "unbundle" 13F institutions to use a more comprehensive dataset of institutional holdings and potentially even households (Gabaix et al., 2022).

We also explore whether asset embeddings explain cross-sectional variation in monthly stock returns. As a point of reference, we use the characteristics of the 5-factor Fama and French model plus the momentum characteristic. This model sets a high bar as these characteristics are known to predict returns and capture significant comovement in returns. We use both unsupervised and supervised asset embeddings. Our initial estimation of asset embeddings using unsupervised methods performs better than simply using the CAPM beta but the unsupervised asset embeddings short of matching the six observed characteristics. However, when using supervised asset embeddings, where we use daily returns from the previous quarter as data to supervise the estimation, asset embeddings explain more of the comovement in returns than the six observed characteristics. The supervised asset embeddings also explain more of the comovement in future returns than six principal components estimated from daily data in the previous quarter. Lastly, we find that the transformer-based AssetBERT model performs significantly better to capture substitution patterns than observed characteristics. We emphasize that these are initial explorations, but they demonstrate the potential of extracting asset embeddings from holdings data.

In Section 6, we discuss several extensions. We show how the basic framework can be extended to extract embeddings when the model in which the embeddings are used is non-linear. We also discuss how to combine observable characteristics (e.g., from accounting data) with asset embeddings. This approach combines the best of both worlds, and can be valuable in practice as Koijen and Yogo (2019) show how traditional characteristics explain some, but not a lot, of the variation in portfolio holdings and returns. Indeed, once aggregated, most of the cross-sectional variation in returns cannot be traced back to observable characteristics. Also, we discuss how this modeling approach can be used to design generative portfolios, which can be viewed as factor-mimicking portfolios that are constructed without using return data. By using salient stocks that have large exposures to a risk factor (e.g., airlines during the COVID-19 pandemic), we can find similar stocks based on their asset embeddings.

Even though our focus in this paper is on asset embeddings, we show how the same models can be used to estimate investor embeddings, which are vector representations of the trading styles of investors. As investor characteristics are broadly limited to institutional type, size, and activeness, investor embeddings can be used to group investors, identify overlap in holdings and crowded trades, and we discuss how investor embeddings can be used for performance measurement.

**Related literature** Our paper relates to various strands in the literature. First, a large literature studies the information contained in observable firm-level characteristics, for instance, to predict returns, model the stochastic discount factor or to predict returns, see Kelly et al. (2019), Kozak et al. (2020), DeMiguel et al. (2020), Feng et al. (2020), Freyberger et al. (2020), Lettau and Pelger (2020); see Kelly and Xiu (2023) for a recent review. A common concern is that the characteristics used are a, potentially small, subset of the information available to investors (Hansen and Richard, 1987). By inverting the asset demand system and extracting asset embeddings from portfolio holdings, we may be able to get closer to the information set used by investors.

Second, a growing literature explores embeddings in the context of text, image, and audio data. Important contributions that are related to our modeling approach are Ducharme et al. (2003), Pennington et al. (2014), Mikolov et al. (2013a), and, for transformer models, Vaswani et al. (2017) and Devlin et al. (2019). in this context of financial market data, Dolphin et al. (2022) estimate a neural network model for stocks using daily returns to capture the covariance between stock returns. Concretely, they define a target stock (say, Apple) and context stocks as those stocks that have a return that is close (in absolute value) to the target stock. With the target and context in hand, they estimate a neural network model analogous to Mikolov et al. (2013a) and explore whether the asset embeddings are useful for hedging purposes.

Third, we are part of a growing literature using machine learning techniques in economics. Early work in asset pricing includes Hassan et al. (2019); Gu et al. (2020); Nagel (2021); Bybee et al. (2023); Chen et al. (2023); Hommel et al. (2023). Related, several recent papers point out the challenges in empirical asset pricing research when characteristics are missing (Bryzgalova et al., 2022; Freyberger et al., 2022). Asset embeddings that are learned from holdings data are ideally positioned to fill this gap.

Fourth, our paper relates to a recent literature on demand system asset pricing, where the main objective is to develop theoretical and empirical frameworks to jointly understand prices, portfolio holdings, and macroeconomic and firm-level variables. Koijen and Yogo (2019) develop a characteristics-based equity demand system that links portfolio holdings to prices, observable characteristics, and residual demand (labeled latent demand). By solving for equilibrium prices, and substituting those in the demand equations, we show that we can "invert" the demand system to learn the characteristics that determine investors' demand.

Lastly, our paper is also related to a recent literature that estimates factor models based on portfolio holdings for various applications, see Madhavan et al. (2021), Gabaix and Koijen (2022), Betermier et al. (2022), and Balasubramaniam et al. (2023). Madhavan et al. (2021) estimate principal components based on ETF and mutual fund holdings to measure trends in crowdedness over time. Gabaix and Koijen (2022) focus on 13F data and sector-level holdings from the flow of funds to isolate idiosyncratic demand shocks, after removing common factors, to estimate demand elasticities using granular instrumental variables (Gabaix and Koijen (2020)). Betermier et al. (2022) develop a factor model of returns that features two new factors beyond the market return. The first factor is sorted by age and the second by wealth. These additional factors explain both holdings and returns. Balasubramaniam et al. (2023) focus on a matrix of stock coholdings. Their main focus is then on uncovering how households tilt their portfolios along the dimension of observable characteristics and clusters of characteristics using factor models. They compare the explained variation of this model to the fraction explained by a latent factor model. Relative to these papers, our main focus is on learning the representation of assets via asset embeddings, and using state-of-the-art AI and ML methods to do so, and to demonstrate their value added in a series of applications. As a by-product, we also learn a representation of investors via investor embeddings.

**Outline** Section 2 introduces asset embeddings after giving a gentle tutorial on natural language processing. Section 3 provides a micro foundation of embeddings data sets with a particular focus on holdings data. Section 4 describes the data. Section 5 contains applications showing how asset embeddings can be used. Section 6 presents extensions, including a programmatic discussion of how our asset and investor embeddings will be useful to address a host of issues in finance. Section 7 concludes.

**Notations** We use the notation $i \in \{1, \dots, I\}$ for investors, $a \in \{1, \dots, A\}$ for assets, and $q$ for quarters. As most calculations happen within a quarter, we often omit subscripts $q$ for simplicity. We normalize a firm's shares outstanding to one. A stock's market capitalization (or price) is denoted by $P_a$, the number shares held by investor $i$ by $Q_{ia}$, and dollar holdings by $H_{ia} = Q_{ia}P_a$. Lowercase letters denote logarithms, e.g. $h_{ia} = \ln H_{ia}$. For a matrix $M \in \mathbb{R}^{I \times A}$ with elements $M_{ia}$, $M_i \in \mathbb{R}^A$ denotes the transpose of $i$-th row; $M_a \in \mathbb{R}^I$ denotes the $a$-th column; $\|M\|_F = \left(\sum_{i,a} M_{ia}^2\right)^{\frac{1}{2}}$ denotes the Frobenius norm of a matrix $M$. For two matrices $B$ and $C$ of the same size, $B \odot C$ is element-wise product matrix, $(B \odot C)_{jk} = B_{jk}C_{jk}$. We use the notation $\beta^e$ for the $e$stimated value of a parameter $\beta$.

# 2 Applying modern AI methods in economics and finance

## 2.1 The main idea: From words and sentences to assets and investors

The field of artificial intelligence (AI) has made rapid progress in recent years in the context of audio, image, natural language processing (NLP) applications. In developing the connection between AI models and economics, we take NLP models as our leading example due to the fitting analogy. The success of those models starts from representing data (e.g., words or speech) as continuous vectors in a potentially high-dimensional space using embeddings. In the context of NLP, embeddings capture the similarity between words, sentiments or documents and can be used to do "math with words." A classic example is to use the vector representations of "king", "woman", and "man" to compute king - man + woman, which yields a vector close to the word "queen".

While we are often interested in modeling the similarity between firms and assets in economics, it is common practice to use observed characteristics. In the AI literature, embeddings are learned from data. This raises the question what data is ideally suited to estimate asset embeddings if the goal is to learn them from data.

Our central insight is a simple, yet broadly applicable one. Just as documents organize words in NLP, images organize pixels in vision, songs organize notes in audio, investors in financial markets organize assets in finance and economics. We therefore modify the modeling frameworks such that we can use holdings data to estimate asset embeddings. In Section 3.1, we provide a simple micro foundation and, importantly, show that holdings data are high-dimensional and can be used to estimate data hungry AI models.

Using this connection, we now discuss at a high, conceptual level how to estimate asset embeddings using NLP models that have been developed during the last three decades using holdings data. This overview is obviously incomplete and our main goal is to explain the analogy between problems and modeling approaches in the NLP literature, and how they can be used to estimate asset embeddings. For textbook treatments of these methods, we refer to Jurafsky and Martin (2023) and Prince (2023).

The traditional approach is Latent Semantic Analysis (LSA) (Dumais et al., 1988), which applies principal components analysis on the document-word matrix. We start by exploring such factors model in Section 2.2 based on the matrix of portfolio holdings across investors for different assets.

The recent AI/ML (machine learning) literature went well beyond those methods. A first important innovation is to use local, non-linear methods, such as Word2Vec ((Mikolov et al., 2013b,a)) and GloVe (Pennington et al., 2014). In case of Word2Vec, embeddings are used in a shallow neural network to predict a target word based on the surrounding words in the target word's context window.[3] To map this model to holdings data, we first rank stocks based on the positions taken by an investor.[4] We then model the probability that an investor holds a position in a given stock, say, Google based on similarly-sized positions in, say, Apple, Microsoft, and Meta using the same neural network. Indeed, just as sentences order words, we exploit the fact that investors order stocks in

[3] Alternatively, we can predict the words in the context of the target word based on the target word itself. The two methods are labeled the continuous bag of words and the continuous skip-gram approach.

[4] Alternatively, we can rank stocks by the deviation of an investors portfolio from the market portfolio (i.e., active holdings) or we can use investors' portfolio rebalancing instead of using the level of holdings..

their portfolios. We discuss this approach in more detail in Section 2.3.[5]

There are two important limitations of the Word2Vec approach. First, the ordering of words within the context window does not matter. Second, the embedding is context-invariant and there is a one-to-one mapping from words to the embedding vectors. These limitations are relevant also in economics: E.g., Apple can be viewed as a technology firm or a large-cap firm, depending on the other firms to which Apple is being compared.

Those limitations are addressed in the third generation of NLP models: transformer models. Recent large language models (LLM) such as BERT and GPT follow this architecture. The key idea is that we start from a context-invariant embedding, and then adjust the embedding by averaging the initial embedding with the embeddings of surrounding words in a sentence or, in our case, surrounding stocks in an investor's portfolio. Hence, instead of a small window of context words, we use the entire sentence or paragraph and multiple layers in a neural network are used to average the initial embeddings to provide context. We provide further details in Section 2.4.

## 2.2 Recommender systems

The simplest thing to do is to estimate embeddings using recommender systems, which are closely connected to LSA models (Dumais et al., 1988). With parameters $\theta = (x_a, \lambda_{iq}, \delta_{iq}, \delta_a, \delta_t, \beta_t)$, we solve:[6]

$$\min_{\theta} \frac{1-\kappa}{N_h \sigma_h^2} \sum_{i,a,q} (h_{iaq} - \delta_{iq} - \delta_a - x_a' \lambda_{iq})^2 + \frac{\kappa}{N_y \sigma_y^2} \sum_{t,a} (y_{at} - \delta_t - \beta_t' x_a)^2, \tag{1}$$

with $N_h$ and $N_y$ the number of non-missing observations of $h_{iaq}$ and $y_{at}$, $\sigma_h^2$ and $\sigma_y^2$ their variances, and $\kappa \in [0,1]$ controls the relative importance of the holdings data and the model data, $y_{at}$. By using the model data, we can extract asset embeddings that account for the task for which the embeddings are eventually used.

The key resulting variables are $x_a$ (asset embeddings) and $\lambda_{iq}$ (investor embeddings), while $\delta_{iq}$, $\delta_a$, and $\delta_t$ are just intercepts. For $\kappa = 0$, this is the core idea of the "recommender systems" that were very successful in the famous Netflix challenge, which asked researcher teams to predict which movies $a$ viewers $i$ would like, given their past ratings $h_{ia}$ (available only for a small subset of movies). Then $x_a$ summarizes the deep characteristics of the movie, and $\lambda_i$ the utility weight of the user on the movie.

It may be worth seeing why embeddings are useful, potentially much more than "observable

[5] GloVe is a related method, which takes the co-occurrence matrix of words across documents as its input, where co-occurring words are measured in a context window around a target word. Next, embeddings are estimated via a weighted principal components analysis, while putting more weight is on words that frequently co-occur. It is straightforward to apply this method in the context of investors' portfolios by forming a matrix of stocks that frequently co-occur in a certain window.

[6] As an extension, one can add penalties to regularize $\theta$, e.g., lasso or ridge. Also, we can put more weight on some investors, e.g., active hedge funds.

characteristics". Suppose that some viewers like teenage vampire movies. It would be very cumbersome for researchers to try to hand-code those characteristics $x_{ak}$, and similar niche characteristics. But a recommender system, estimating $x_a$ without hand-coding, will automatically generate characteristics $k$ that are akin to "teenage" and "vampire", and will give movies of that genre a high $x_{ak}$; and their aficionados will have a high $\lambda_{ik}$ for those characteristics.

Likewise, suppose that some investors like stocks with "high number of clicks received by the firm's web site", at least point in time (e.g. late 1990s). This information is not in plain accounting data, and would be hard to encode. However, a recommender system (and more generally, ML techniques) will uncover it, create a characteristics $x_{ak}$ akin to "number of clicks" for firm $a$, and the investors paying attention to it will have a high $\lambda_{ik}$.

We next discuss implementation specifics. As not all investors hold all stocks, we implement two variants. First, we complete the panel by replacing missing values by zeros (after removing a stock and investor-time fixed effect). Second, we optimize (1) directly over only the non-missing values.[7] Also, as the remaining models in this section only use the ordering of stocks in an investor's portfolio, we also estimate the model using the percentile rank of an investor's holdings. By comparing the asset embeddings in levels and in terms of percentile ranks, we can assess the loss in information when using ranks instead of level holdings.

Equation (1) can be estimated on a single quarter of data or multiple quarters. In case of multiple quarters, (1) embeds the economic assumption that asset embeddings $x_a$ are stable over the estimation period, while investor embeddings, $\lambda_{iq}$, fluctuate. Intuitively, $x_a$ captures a particular trading direction (e.g., buy Zoom and short airlines during the COVID-19 pandemic) and $\lambda_{iq}$ the investor's propensity to trade in that direction (which may change from period-to-period as the investors' views change).

To estimate the model, we can use $h_{ia}$ or $\tilde{h}_{ia} = \ln \frac{H_{ia}/H_i}{H_a/H_M}$, which are the "normalized holdings" or "active weights", where $H_{ia}/H_i$ is the fraction of investor $i$'s assets invested in asset $a$, and $H_a/H_M$ is the fraction of the market in asset $a$. Alternatively, we can use changes in holdings or active holdings.

## 2.3 Simple neural network models: Word2Vec

The next important step in the NLP literature is Word2Vec (Mikolov et al. (2013a,b)). While PCA considers all the words and tries to fit a global model, Word2Vec pays attention to words in a narrow window around a target word.

To estimate the model, we take advantage of the fact that investors order assets just as sentences order words. To apply the Word2Vec model, we therefore compute the rank, $\rho_{ia}$, of asset $a$ in investor $i$'s portfolio. For a given investor $i$, we then try to predict the name ranked $\rho$, given that we are

[7] To solve for asset and investor embeddings numerically, we can use alternating least squares.

given the names of the neighboring stocks.[8] Call $\mathcal{N}_\rho = \{b : 0 < |\rho_i - \rho_{ib}| \leq \delta\}$ the known assets in neighborhood of rank $\rho$, i.e., the stocks whose rank differs by at most $\delta$ or rank $\rho$; and define $\bar{x}_{i\rho} := \frac{\sum_{b \in \mathcal{N}_{i\rho}} x_b}{|\mathcal{N}_{i\rho}|}$ to the average value of the embedding around rank $\rho$ for investor $i$. Intuitively, the missing stock $a$ at rank $\rho$ can be expected to have similar characteristics to those of neighboring stocks, so close to $\bar{x}_{i\rho}$. Word2Vec uses this idea, and uses $\bar{x}_{i\rho}$ to predict the asset $a$ at rank $\rho$, with in the following manner:

$$\mathbb{P}(\rho_{ai} = \rho \mid \bar{x}_{i\rho}) = \frac{\exp(x_a' \bar{x}_{i\rho})}{\sum_c \exp(x_c' \bar{x}_{i\rho})} \tag{2}$$

i.e., $x_a$ should have high covariance with $\bar{x}_\rho$. The embeddings are estimated to maximize that likelihood.

The methodology discussed so far is designed to provide asset embeddings. We can follow the same logic to construct investor embeddings. Instead if ranking stocks for a given investor, we can rank investors for a given stock. We can then predict the probability that investor $i$ is the $n$-th largest owner of a stock based on the other investors holding the stock.

## 2.4 Transformer models

In the models discussed so far, there is a one-to-one mapping from a stock to an embedding vector, $x_a$. The recent explosion of work in the NLP literature uses transformer models introduced by Vaswani et al. (2017) that are central to recent large language models including BERT[9] models (Devlin et al., 2019) and generative pre-trained transformer (GPT) models (Radford et al., 2018). Transformer models are deep neural network models that add attention and result in contextualized embeddings. Transformer models can feature an encoder and/or a decoder part. In case of BERT, which is the modeling architecture we adopt in the paper, the model features only the encoder layers of the transformer.[10]

A key question is how this modeling architecture can be used with *numerical* values instead of discrete *words*. That is, how to transform numerical information on holdings $h_{ia}$ into BERT-readable language, which uses words. In this section, we propose a solution to this problem and refer to the resulting BERT model estimated using numerical holdings data as AssetBERT.[11]

[8] Here we describe the "continuous bag of words" version of Word2Vec. There is also a "continuous skip-gram" version, which predicts the whole neighborhood with just one word.

[9] BERT stands for Bidirectional Encoder Representations from Transformers.

[10] The GPT architecture uses constrained self-attention using only the context to the left so it can be used for text generation. The BERT architecture, as we discuss below in more detail, is bidirectional and uses information from the left and the right of a word. In the context of portfolio holdings, economic intuition suggests that using information from the entire portfolio is most meaningful to obtain contextualized asset embeddings. However, as we will discuss in Section 6.8, the model can be used to generate portfolios and it may be the case that GPT-style architectures are particularly well suited for that.

[11] We have not seen this solution in the vast BERT literature, but we do not know it all. We welcome pointers to antecedents of our proposed solution.

### 2.4.1 Main intuition and a simple example of attention

The basic idea can be understood in a simple example. There are numerous articles and blog posts explaining this intuition, including textbook treatments in Jurafsky and Martin (2023) and Prince (2023), and we specialize it to our context. Suppose that the set of stocks held by investor $i$ is $\mathcal{N}_i$, and we are interested in the embedding of asset $a \in \mathcal{N}_i$, which here is a stock. We start from a stock's initial embedding $x_a$ (the "query"), and compute the similarity score with all stocks in the investor's portfolio, $x_b$ (the "keys")

$$\sigma_{ab} = x_a' x_b.$$

We then compute the contextualized embedding, $x_a^i$, by computing the weighted average of the embeddings in the investor's portfolio

$$x_a^i = \sum_{b \in \mathcal{N}_i} \frac{e^{\sigma_{ab}}}{\sum_{c \in \mathcal{N}_i} e^{\sigma_{ac}}} x_b.$$

The embedding used in the weighted average, $x_b$, is referred to as the "value." Intuitively, if $x_a$ is an embedding vector for an asset capturing information about a firm's industry, reliance on external finance, and supply-chain risk. Then, depending on the other stocks in an investor's portfolio, we obtain a different embedding vector for the same firm, perhaps better reflecting the firm's industry or external financing conditions.

In computing the contextualized embedding, we used the original embedding, $x_a$, in three places, namely for the query, keys, and values. The transformer model then generalizes this logic and uses $q_a = W^Q x_a$ as query, $k_a = W^K x_a$ as keys, and $v_a = W^V x_a$ as values, where the $W$ are matrices to be estimated. The embedding is then computed as

$$x_a^i = \sum_{b \in \mathcal{N}_i} \frac{e^{\sigma_{ab}}}{\sum_{c \in \mathcal{N}_i} e^{\sigma_{ac}}} v_b, \qquad \sigma_{ab} = q_a' k_b.$$

As is clear from this structure, the attention mechanism takes into account an investor's entire portfolio. The transformer encoder then stacks multiple of such layers and adds a feedforward layer after each attention layer. Moreover, instead of a single attention layer, there are multiple attention heads that help to speed up the calculations and may improve the model's performance as the multiple attention heads can focus on different aspects of an investor's portfolio. Michel et al. (2019) provides an in-depth discussion on the benefits of multi-headed attention mechanisms.

What is still missing from the model so far is the location of the words or, in our case, the position of the stock in an investor's portfolio. Without such information, the model treats the inputs as a bag of words (or bag of stocks). To incorporate this important information, we rank all stocks in an investor's portfolio based on the holdings (or, alternatively, active weights or rebalancing) to determine a stock's position. We then add position embeddings to the initial asset embeddings

$x_a$ and use this instead as inputs into the model. Just as the initial asset embeddings and the parameters of the attention layers, the position embeddings are estimated in BERT models.

### 2.4.2 AssetBERT: Architecture choices and inference

Before explaining our AssetBERT's architecture and how we estimate it, we summarize the classic BERT model and how it has been estimated. The original BERT model has been estimated based on two tasks: masked language modeling and next-sentence prediction. As next sentence prediction is not directly relevant in our setting, we focus on masked language modeling only. In case of masked language modeling, one (or multiple) of the input tokens is masked in a sentence. The model then generates a prediction for the masked word based on the final embedding for that word (which depends on the other words in the sentence via the attention mechanism). Indeed, the model generates a distribution over words for the masked word based on the embedding after the final transformer block. In practice, 15% of the words are masked.[12] The model is then estimated to maximize the probability of picking the right word on a training sample and evaluated on a test sample to make sure that the model does not overfit.

To estimate AssetBERT, our baseline solution is the following. For each investor $i$, we order the assets' names $a^i(1), a^i(2), \ldots, a^i(A)$, by decreasing holdings sizes $h_{ia}$, so that $a^i(k)$ is the name of the $k$-th largest position of investor $i$. This is a "sentence," which would look like, for a hypothetical investor 1:[13]

$$\text{Apple, IBM, Tesla,...., Walmart} \tag{3}$$

We start with a sentence or investor 1, and then have a second sentence for investor 2, which might be IBM, Cisco, Apple,..., Ford. The tokens are just the stock names, without further tokenization (except for e.g. an end-of-sentence separator token, like [CLS] and [SEP]).[14] So, there are $A$ tokens, the number of assets, plus a few extra BERT-specific tokens.

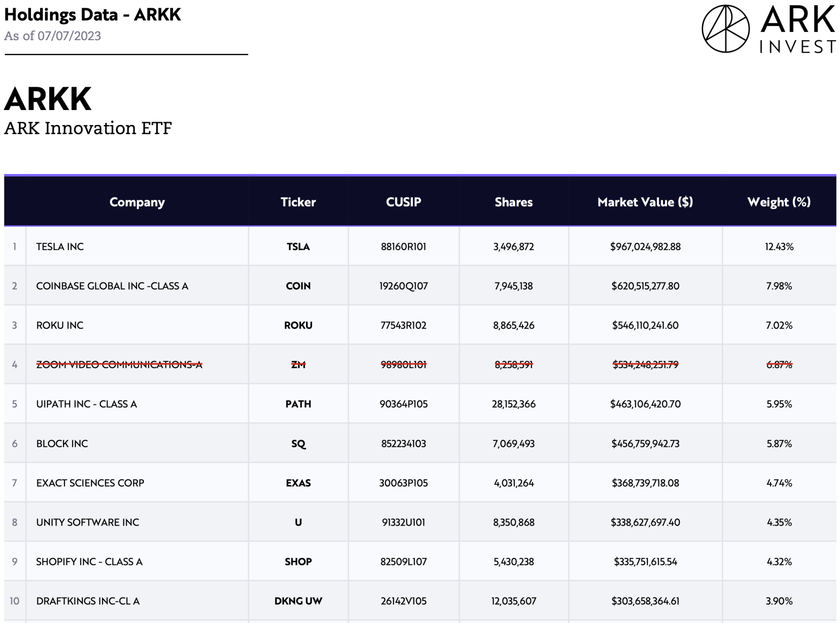
To estimate the model, we mask one or more of the stocks. See for instance the Ark exchange traded fund (ETF) in July 2023 in Figure 1 (Ark was then a popular technology fund). In this example, the model sees the holdings of all stocks in the Artk ETF, except the one at holdings rank 4, which turns out to be Zoom (in practice, we mask more than one stocks). The task for the model is to predict the identity of that asset at rank 4. The parameters are then chosen the maximize the likelihood that the model predicts the correct stock, for this task and similar masked assets at other institutions. Technically, the metric of performance is cross-entropy.

[12] Of the masked tokens, 80% receive the token [MASK], 10% are the actual token, and 10% are replaced by a random token, see Appendix A of Devlin et al. (2019). We follow the same masking scheme for AssetBERT.

[13] Because a sentence can have a maximal length, we actually break it in smaller chunks, but this is conceptually not important.

[14] There are five BERT special tokens, [CLS] for the beginning of a sentence, [SEP] for the end of a sentence, [UNK] for unknown tokens (which does not happen in holdings data), [PAD] (to complete the length of the sentence to a standard batch size of, in our case, 64), and [MASK] (for the masked tokens).

Figure 1: **Example of prediction task for masked holdings** The model sees the holdings of all stocks in the Ark ETF, except the one with holdings rank 4. The task is to guess (i.e., form a probability distribution) the identity of this masked stock, which turns out to be Zoom. This is analogous to the situation in language models, where a typical task is to predict a missing word in a sentence like “Please pass me the _____ and pepper”.



**Holdings Data - ARKK**
As of 07/07/2023

ARK INVEST

**ARKK**
ARK Innovation ETF

| | Company | Ticker | CUSIP | Shares | Market Value ($) | Weight (%) |
|---|---|---|---|---|---|---|
| 1 | TESLA INC | **TSLA** | 88160R101 | 3,496,872 | $967,024,982.88 | 12.43% |
| 2 | COINBASE GLOBAL INC -CLASS A | **COIN** | 19260Q107 | 7,945,138 | $620,515,277.80 | 7.98% |
| 3 | ROKU INC | **ROKU** | 77543R102 | 8,865,426 | $546,110,241.60 | 7.02% |
| 4 | ~~ZOOM VIDEO COMMUNICATIONS-A~~ | ~~**ZM**~~ | ~~98980L101~~ | ~~8,258,591~~ | ~~$534,248,251.79~~ | ~~6.87%~~ |
| 5 | UIPATH INC - CLASS A | **PATH** | 90364P105 | 28,152,366 | $463,106,420.70 | 5.95% |
| 6 | BLOCK INC | **SQ** | 852234103 | 7,069,493 | $456,759,942.73 | 5.87% |
| 7 | EXACT SCIENCES CORP | **EXAS** | 30063P105 | 4,031,264 | $368,739,718.08 | 4.74% |
| 8 | UNITY SOFTWARE INC | **U** | 91332U101 | 8,350,868 | $338,627,697.40 | 4.35% |
| 9 | SHOPIFY INC - CLASS A | **SHOP** | 82509L107 | 5,430,238 | $335,751,615.54 | 4.32% |
| 10 | DRAFTKINGS INC-CL A | **DKNG UW** | 26142V105 | 12,035,607 | $303,658,364.61 | 3.90% |

### 2.4.3 AssetBERT: Extensions in progress

We discuss several extensions that we are exploring in ongoing work, including in future iterations of this paper.

**Time-varying asset embeddings** In our current implementation of these models, we focus on a single quarter, or several quarters, to estimate the models. These rolling estimates are simple to implement, but they are likely inefficient. Also, this problem is more unique to our setting compared to language models. While some aspects of a firm or asset are highly persistent over time, others are more likely to change over time (e.g., profitability and riskiness). While holdings reflect this new reality of firms, we may still want to use data from older quarters as well. We discuss two different approaches to model time variation in the context of transformer models.[15]

First, the model is estimated by masking 15% of the stocks in the sample. Instead of applying the masking probability equally across quarters, we can lower the masking probability in earlier quarters in the sample. This implicitly puts less weight on older data. Second, we can estimate the model in two steps. In this case, we use a longer sample (e.g., 5 or 10 years of data) to estimate the

[15] For recommender systems or simple neural network models, there are various intuitive approaches to introduce dynamics, such as putting different importance weights on different terms in (1) or by applying the insights from generalized autoregressive score models (Creal et al., 2013).

model, and we then fine-tune the model on the current quarter or the last couple of quarters. In this case, we use a long history to estimate the longer-run similarities between firms and the more recent data to update those estimates.

**Integrated asset and investor embeddings** There are other solutions that offer useful variants. One is to use tokens for both assets and investors. Call $I_i$ the token for investor $i$. Then, the sentence corresponding to investor $i$ is $I_i, a^i(1), I_i, a^i(2), \ldots, I_i, a^i(A)$. For instance the first sentence is:

$$I_1\text{, Apple, } I_1\text{, IBM, } I_1\text{, Tesla,...., } I_1\text{, Walmart} \tag{4}$$

This way, BERT learns simultaneously both an asset embedding $x_a$ for asset $a$, and and an investor embedding $\lambda_i$ for investor $i$.[16] There are $A+I$ tokens, plus the few BERT-specific tokens like [CLS] and [SEP].

Yet another solution would be to run two BERTs, and glue them. First, we run the basic BERT with $a^i(1), a^i(2), \ldots, a^i(A)$. This yields asset embeddings $x_a$. Second, we run another BERT flipping the roles of investors and assets: for each asset $a$, a sentences is $I^a_{(1)}, \ldots, I^a_{(I)}$,where $I^a_{(k)}$ is the token of the $k$-th largest investor of asset $a$ (perhaps in normalized holdings space). That gives an investor embedding $\lambda_i$ for each investor $i$. When we use those asset embeddings $x_a$ and investor embeddings $\lambda_i$, via a third estimated neural network, to predict holdings $h_{ia}$. In what follows, we present results for the basic "numbers to BERT" solution, but we plan to explore other variants in future iterations of this paper.

**Under- and over-weights in investor portfolios** One unique aspect of holdings data compared to text is that holdings that are far apart in an investor's portfolio may actually be tightly connected. For instance, an investor concerned about the COVID-19 pandemic may overweight technology companies and underweight airlines, both relative to the market portfolio. In sentences, paragraphs or documents, such a tight link is often less obvious. While the model may be able to discover such subtleties via position embeddings, we may be able to capture this more directly by focusing on active holdings, $\tilde{h}_{ia} = \ln \frac{H_{ia}/H_i}{H_a/H_M}$. Active holdings are the "normalized holdings" or "active weights", where $H_{ia}/H_i$ is the fraction of investor $i$'s assets invested in asset $a$, and $H_a/H_M$ is the fraction of the market in asset $a$. Instead of ranking stocks based on $\tilde{h}_{ia}$, we rank them based on $\mid \tilde{h}_{ia} \mid$ and make a duplicate of every stock $a$, $a^+$ when $\tilde{h}_{ia} \geq 0$ and $a^-$ otherwise. We can then find for stock $a$ and $b$ that $a^+$ and $b^-$ have similar embeddings, implying that they are opposite bets of the same trade.

[16] One related advantage of this protocol is that BERT will process, among other things, non-linear functions $h(x_a, x_i)$, which seems useful to describe investors' decisions.

### 2.5 Investor embeddings

We have focused throughout this section on asset embeddings, and only discussed investor embeddings in passing. However, as all methods can be used to estimate investor embeddings as well, we briefly discuss two applications of those in finance applications. We leave a detailed investigation of those to future work however.

First, investor embeddings can be used to discover crowded trades as investor embeddings can be used to cluster investors in terms of holdings or trading behavior. Second, investor embeddings can be used for performance measurement. Building on the insights of Daniel et al. (1997), who develop a characteristics-based benchmark, our investor embeddings can be used to do this in higher dimensions. An investor $i$ with investor embedding $\lambda_i$ can be compared to investors with embedding vectors $\lambda_j$ such that $\|\lambda_i - \lambda_j\|_F \leq \delta$, and we can compute $\alpha_{it} = r_{it} - \frac{1}{N_{it}} \sum_{j:\|\lambda_i - \lambda_j\|_F \leq \delta} r_{jt}$ with $N_{it}$ the number of investors that are sufficiently close to investor $i$.

## 3 Portfolio holdings as embedding data

A key feature of embedding data is that there are two dimensions in a given period, say, a quarter. For instance, we can use daily returns across stocks, daily volume data or word distributions of firms' 10-K filings. We argue in this section that holdings data, and transformations thereof, are particularly well suited as embedding data. We start with a simple equilibrium model of asset demand in Section 3.1 to micro-found the use of holdings data as embedding data. Appendix A provides additional derivations.

### 3.1 Theoretical motivation

We model log dollar demand of investor $i$ for asset $a$ as

$$h_{ia} = c_i^h + (1 - \zeta_i) p_a + \nu_{ia}, \tag{5}$$

where $\zeta_i$ is the investor's demand elasticity, $\nu_{ia}$ the demand shifter, and $c_n^h$ captures the investor's size.[17] We assume that the demand shifter has a factor structure,

$$\nu_{ia} = x_a' \lambda_i^\nu + u_{ia}, \tag{6}$$

where $x_a$ is the *asset embedding* and it captures how much stock $a$ is affected by each embedding dimension. The *investor embedding*, $\lambda_i^\nu$, captures how an investor tilts its portfolio to the various

[17] In the language of Gabaix and Koijen (2022) this $\zeta_i$ is the micro elasticity $\zeta_i^\perp$ (price elasticity of demand when choosing between different stocks), rather than the macro elasticity (price elasticity of demand when choosing between the aggregate stock market vs the aggregate bond market), but for notational simplicity we omit the $\perp$.

embedding dimensions. In the simple model of this section, they enter linearly, but they can enter nonlinearly in more general models (e.g. with bounds that prevent an investor from over- or under-weighting a stock by too much). The residuals, $u_{ia}$, are idiosyncratic bets of investor $i$ on stock $a$. There are different micro foundations that give rise to this model of demand (Koijen and Yogo, 2019; Koijen et al., 2022), for instance by assuming that investors have mean-variance preferences and that returns follow a factor model. The asset embeddings, $x_a$, then capture variation in expected returns (or alphas) across assets and heterogeneity in factor loadings.

In Appendix A, we show that when we impose market clearing, solve for asset prices, and substitute the equilibrium prices in (5), we obtain the equilibrium holdings

$$h_{ia} = \phi_i^h + \phi_a^h + \lambda_i' x_a + \epsilon_{ia},$$

where $\lambda_i$ is an linear transformation of $\lambda_i^\nu$ and $\zeta_i$. Only when all investors are identical in terms of $\lambda_i^\nu = \lambda^\nu$ and $\zeta_i = \zeta$, then $\lambda_i = 0$ and we cannot use equilibrium holdings to recover asset embeddings. However, this is an empirically irrelevant special case given the heterogeneity in investment strategies and trading behavior observed empirically.

We also show in Appendix A that when asset embeddings do not change too much over time, $\Delta x_{aq} \simeq 0$, then we have for rebalancing

$$\Delta h_{iaq} = \Delta\phi_{iq}^h + \Delta\phi_{aq}^h + \Delta\lambda_{iq}' x_{a,q-1} + \Delta\epsilon_{iaq},$$

and for returns

$$r_{aq} = \frac{1}{\zeta_S} \left(\Delta\lambda_{Sq}^\nu\right)' x_{a,q-1} + \Delta u_{Saq},$$

where $X_S \coloneqq \sum_i S_i X_i$ for importance weights of investors $S_i$, $\sum_i S_i = 1$.

## 3.2 Do we have enough data to estimate asset embeddings?

One obvious concern is that large language models are high-dimensional and flexible, and therefore require vast amounts of data in estimation. We compare the relative proportions of data, parameters and "words" (i.e., in our setting, assets or investors) in machine learning compared to our financial setting. Those ratio are of course just indicative, rather than guarantees of success, but we found it useful to consider them. The punchline is the following. In LLMs, typically, the ratio of the number of data points relative to number of tokens (or vocabulary size) is around $10^4 - 10^5$. In our financial data, it is around $10^4$. So this gives some a priori hope that AI/ML models can be used to estimate asset embeddings.

Let us now detail the analysis, starting with LLMs. For concreteness, in the base BERT model, the size of the input embeddings (which determines the size of the hidden layers) is 768. There are then 12 layers of transformers of transformer blocks, where each layer has 12 attention heads. The total number of parameters is 110 million. The large BERT model (featuring 24 attention layers, an embedding size of 1024, and 16 attention heads per layer) has 340 million parameters. The number of unique tokens is 30,000. The model is trained on 3.3 billion words (a combination of the English Wikipedia, containing 2.5 billion words, and the Toronto Book Corpus, containing 800 million words). In case of Google's Word2Vec (Mikolov et al., 2013a), a training data set (Google news data) of one billion words was used with a vocabulary of 692 thousand. Mikolov et al. (2013b) contains other data sets of various sizes, ranging up to 6 billion words. Hence, the ratio of data to tokens or vocabulary is around $10^4 - 10^5$.

How does this align with data from financial markets? Before discussing the size of the data, we make three observations. First, the dimensionality of the embedding vector (e.g., 768 for BERT) is very high in language models, and much larger than the number of characteristics that we typically consider. It seems therefore natural to start with smaller models and then see whether scaling them up adds much benefits. Second, there is a growing awareness that data quality is important in estimating the model (Gunasekar et al., 2023). While text data tends to be messy, and there is vast literature on how to tokenize text, our holdings data are from regulatory filings and contain hardly any measurement error. Tokenization is also trivial as we directly observe the stock identifiers. Third, the number of tokens (30,000) is much larger than the number of firms. For these three reasons, we can potentially estimate smaller-scale models and require less data relative to the number of parameters as the data are of higher quality.

As is clear from the earlier discussion, we need two dimensions to estimate the model. For instance, in a quarter or year, we can use daily data on returns and volume, or we can use portfolio holdings across investors. Of course, one can also combine these data sources. We focus on holdings data due to its high dimensionality.

We provide three examples. First, we use quarterly filings of Securities and Exchange Commission Form 13F from FactSet (Section 4 discusses the data in more detail). All institutional investment managers that exercise investment discretion on accounts holding Section 13(f) securities, exceeding \$100 million in total market value, must file the form. From 2000.Q1 to 2021.Q4, the data contain over 56 million investor positions for fewer than 10,000 unique 13(f) securities. Second, 13F data is aggregated at the institutional level; for instance, Vanguard counts as a single institution and we do not have separate holdings for their sector funds, style funds, et cetera.[18] There are also detailed filings for open-end mutual funds, closed-end funds, variable annuity funds, and exchange-traded (ETF) funds. During the same period and using quarterly data, there are a little over 50 million positions. Third, in addition to institutional holdings data, there are also detailed data on households' portfolios in the United States from Addepar (Gabaix et al., 2022). While these data only start in 2016, the sample is rapidly expanding with over 150,000 investors in

[18] In some cases, we have multiple filings per institution but they are typically at a fairly high level of aggregation.

2021.[19]

Taken together, we have around $10^4$ observations per security This ratio improves for larger firms. In addition, the models that we are interested in estimating are much smaller. A main drawback that we face, however, is that asset embeddings can change more rapidly over time, as discussed in Section 2.4.3. Hence, the performance of these models ultimately is an empirical question, to which we turn next.

# 4 Data

## 4.1 Data sources and sample selection

We combine data on equity prices from CRSP, accounting data from Compustat, and holdings data from FactSet. The data construction follows Koijen and Yogo (2019) and Koijen et al. (2022). The sample period is from 2000.Q1 to 2020.Q4. We define the change in the number of shares as $\Delta q_{ia} = 2\frac{Q_{ia}-Q_{ia}^{-}}{Q_{ia}+Q_{ia}^{-}}$, where $Q_{ia}^{-}$ denote the number of shares held in the previous quarter. This definition is less prone to outliers and ensures that $\Delta q_{ia} \in [-2, 2]$.

Our main focus for holdings data is on 13F filings and we confirm the robustness of our results using data from mutual funds, ETFs, closed-end funds, and variable annuity funds. In future versions of the paper, we will “unbundle” 13F institutions when possible to use the most granular data available in terms of institutional holdings.[20]

In selecting our sample, we remove micro caps as measured by the 20th percentile of the size distribution of all stocks listed at the NYSE. We keep stocks that are held by at least 20 investors and investors who hold at least 20 stocks. When using historical data on holdings beyond the current quarter, we restrict attention to the list of stocks in the current quarter as the “vocabulary.”

## 4.2 Empirical variants and a benchmark

The asset embeddings that we extract differ in terms of the embedding data sets that we use, in terms of the history of data, in terms of the methodology. For the embedding data sets, we use, alternatively: log holdings, $h_{ia}$; log active weights, $\tilde{h}_{ia}$; rebalancing, $\Delta q_{ia}$.

To provide a point of reference for the asset embeddings that we estimate, we use the characteristics from the asset demand system in Koijen and Yogo (2019) for valuation (which are log book equity, profitability, asset growth, beta, and dividends-to-book equity) and the 5-factor Fama and French characteristics plus momentum for return co-movement. The details of the construction are

[19] The ideal data for this study would be from custodians, who observe portfolios from a large number of investors.

[20] It is possible to further expand the holdings data by using data on bond mutual funds that hold the corporate bonds of the same firms.

discussed in Koijen and Yogo (2019). While these lists of characteristics are by no means exhaustive, they provide a reasonable point of reference to benchmark the estimated asset embeddings.

# 5 Applications

We explore three empirical applications to demonstrate how asset embeddings can be estimated and used. We discuss a series of additional applications that we plan to explore in future versions of this paper in Section 6. In Section 5.1, we explore how embeddings can be used to determine a firm's valuation. In Section L.2, we study the risk in stock returns by measuring co-movement. In Section L.3, we show how asset embeddings can be used to estimate substitution patterns. We stress that this part of the paper is very much work in progress, and we are actively exploring modeling variants and extending the scope of the data.

## 5.1 Stock valuations

In our first application, we use asset embedding to determine a firm's valuation (see Hommel et al. (2023) for a recent comparison of corporate valuation methods).

### 5.1.1 Framework and implementation

We first construct a valuation residual, $m_a^{\perp}$, which is the residual from a regression of a firm's log market capitalization, $m_a$, on its log book equity, $be_a$,

$$m_{aq} = \gamma_{0q} + \gamma_{1q} be_a + m_{aq}^{\perp}, \tag{7}$$

which is implemented using separate cross-sectional regressions for each quarter. We then implement two exercises:

1. Explaining valuations.

   In this simple first exercise, we measure the explanatory power of observed and estimated asset embeddings for the cross-section of valuation residuals

   $$m_{aq}^{\perp} = \beta_{0q} + \beta_{1q}' x_{aq} + \epsilon_{aq}, \tag{8}$$

   and record the $R^2$, $R_q^2$. We then report the average value across quarters, $\frac{1}{Q}\sum_q R_q^2$.

2. Predicting valuations out of sample.

   We split the sample in a estimation and test sample based on a 80/20 split. We then estimate (8) for the estimation sample, and record the estimates, $(\beta_{0q}^{e}, \beta_{1q}^{e'})'$. We compute the out-of-sample $R_q^{2,OOS}$ as

$$R_q^{2,OOS} = 1 - \frac{Var_q(m_{aq}^{\perp} - \beta_{0q}^{e} - \beta_{1q}^{e\prime} x_{aq})}{Var_q(m_{aq}^{\perp})},$$

for the data in the test sample. We report the average out-of-sample $R^2$ values, $\frac{1}{Q}\sum_q R_q^{2,OOS}$.

The second exercise measures whether we overfit the model and can be used as an alternative for the traditional comparables analysis, we discuss in more detail below.

Before seeing the results, it is worth pondering what we should expect given the model's logic. Suppose that (i) the bilinear model with $K$ in (3.1) is fully correctly specified, and (ii) the $\lambda_i$ span the $K$ dimensions (precisely $\sum_i \lambda_i \lambda_i'$ has full rank $K$). Then as the number of investors $I \to \infty$, one should be able to fully recover the $x_a$.[21] If, in addition, (iii) the estimation sample is very large ($A \times 0.9 \to \infty$), then $\beta_1^e$ should be estimated without error, and the $R_{AE}^2$ should be 1.

This simple observation also gives a guide to what might go wrong, and how to improve the procedure. For instance, the $R^2$ will be much less than 1 if the bilinear specification (6) is importantly incorrect, e.g. because $\nu_{ia}$ is non-linear function of $x_a$ and $\lambda_i^{\nu}$ (for instance, the investor might not want take a very large position in any given stock). Then a non-linear specification of embeddings and the valuation $m_a^{\perp} = f\left(\beta_1, x_a^e\right)$ would be in order, and this can be explored in the context of the transformer model.

### 5.1.2 Empirical results

We report the main results in Figure 2. We report the in-sample fit in the left panel and the out-of-sample fit in the right panel. The patterns are very similar across the two panels. The first two bars correspond to observable characteristics. The first bar only uses the CAPM beta, while the second bar uses the characteristics of the asset demand system of (Koijen and Yogo, 2019). For the next two bars, we estimate asset embeddings of the same dimension, $K = 5$, using PCA by replacing the missing values in the panel by zeros (after de-meaning the data). In the fourth bar, we use percentile ranks as opposed to log holdings. The main takeaway is that the estimated asset embeddings perform significantly better, both in- and out-of sample.

The next two bars follow the same structure but we use the recommender system to estimate the asset embeddings. This implies that we do not have to replace the missing values by zeros. The recommender system further improves the performance of asset embeddings for this task. We then explore the Word2Vec model with a context window of 5 stocks on either side of the target stock.[22] Word2Vec performs similarly to the PCA model with missing values replaced by zeros. We report the performance of "placebo" embeddings (constructed by simulating $x_a$ from a standard normal distribution) in the final column, which perform poorly, both in- and out-of-sample. This

[21] These statements come from the basic results from PCA.

[22] It is possible to let the size of the context window depend on the size of an investor's portfolio.

Figure 2: **Using asset embeddings to explain firm valuations** We compare the in-sample (left panel) and out-of-sample (right panel) performance of asset embeddings estimated using various methods based on log holdings. The bar "Beta" uses the CAPM beta, the bar "Observed" uses the characteristics in Koijen and Yogo (2019), "PCA0" uses PCA where the missing values are replaced by zeros, "PCA0-R" uses PCA, replaces the missing values by zeros, and replaces holdings by percentile ranks, "RecSys" uses the recommender system, "RecSys-R" uses the recommender system and replaces holdings by ranks, "W2V" uses the Word2Vec model, and "Placebo" are randomly-simulated embeddings. All models are estimated with the same dimensionality as the observed data ($K = 5$) and we use a single quarter of data. The reported $R^2$ is an average of the quarterly cross-sectional $R^2$ values. The sample is from 2005.Q1 to 2020.Q4 and uses holdings data from 13F filings.

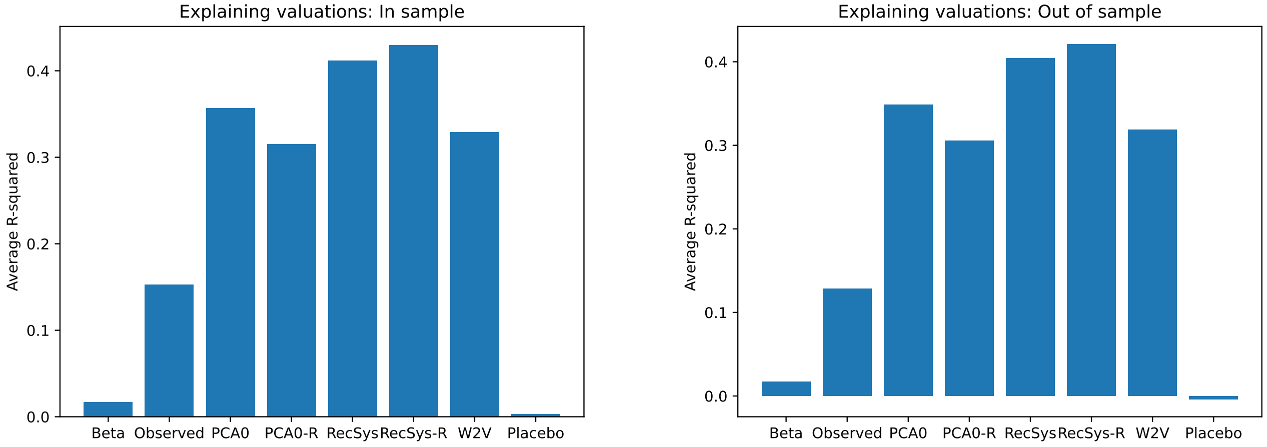
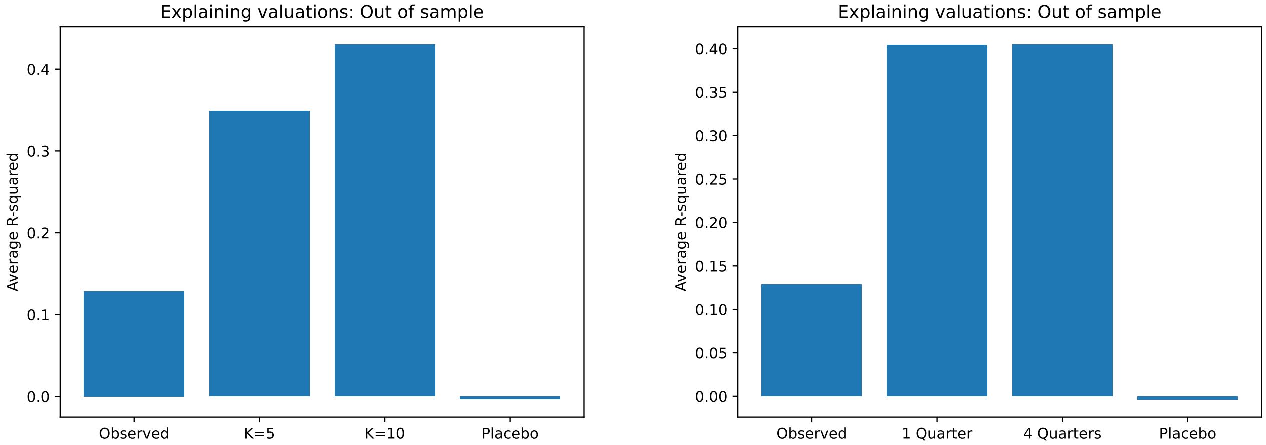


Figure 3: **Using asset embeddings to explain firm valuations: Depth of the embedding and sample size** We compare out-of-sample performance of asset embeddings estimated using recommender systems. In the left panel, we use embeddings of depth $K = 5$ or $K = 10$. In the right panel, we use $K = 5$, but use 4 quarters of data instead of a single quarter. The reported $R^2$ is an average of the quarterly cross-sectional $R^2$ values. The sample is from 2005.Q1 to 2020.Q4 and uses holdings data from 13F filings.



Explaining valuations: Out of sample

Explaining valuations: Out of sample

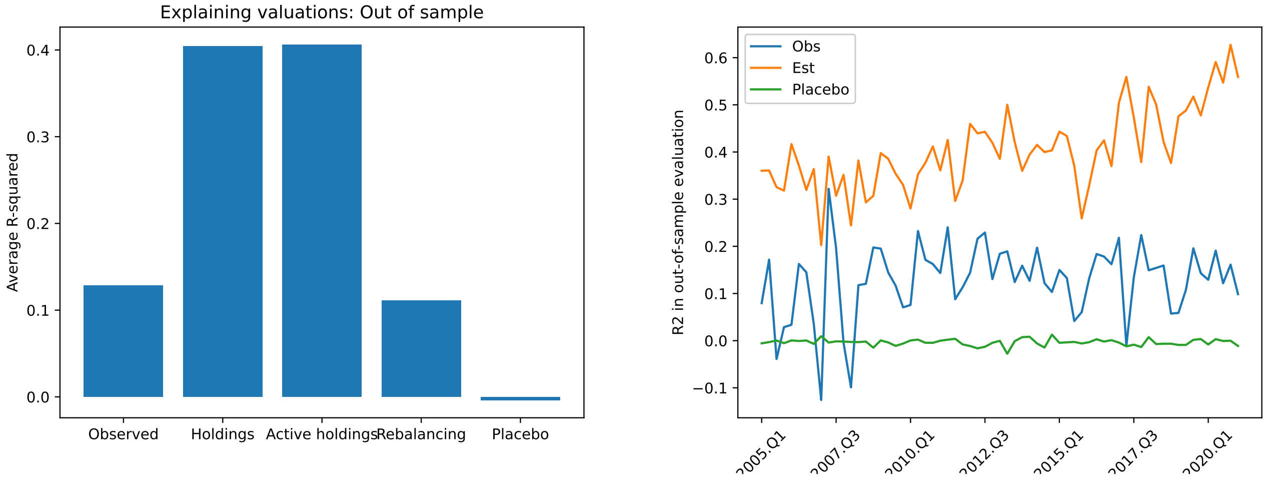
implies that the performance improvements are due to additional information in the estimated asset embeddings compared to observable characteristics.

We explore the sensitivity of our findings to changing the depth of the embeddings from $K = 5$ to $K = 10$ in the left panel of Figure 3. We focus on the recommender system to estimate the asset embeddings. While the performance naturally improves in sample, the performance still improves out of sample. That said, we do not approach an $R^2$ close to 100%, implying that we are still missing important information in valuations that cannot be extracted by the linear recommender system. In the right panel, we use either one quarter of four quarters of data. We find that the performance is very similar when we use a longer sample.

Next, we we explore using holdings, active holdings, or rebalancing in the left panel of Figure 4. While using log holdings and active holdings are very similar (as is to be expected given the fixed effects), the results deteriorate when we use data on portfolio rebalancing. In the right panel, we show the stability of the $R^2$ over time. Quite interestingly, the explained variation by asset embeddings seems to trend up during our sample period. The explained variation by observable characteristics is, by contrast, stable.

Lastly, in Figure 5, we report the results for fund holdings (mutual funds, exchange-traded funds, closed-end funds, and variable annuity funds) instead of 13F filings. As before, the left panel contains the in-sample results and the right panel contains the out-of-sample results. We find that the results strengthen somewhat when using the fund-level data, and the recommender system now recovers asset embeddings that explain around half of all variation in the valuation residual. In

Figure 4: **Using asset embeddings to explain firm valuations: Embedding data and stability over time** We compare out-of-sample performance of asset embeddings estimated using recommender systems. In the left panel, we use embeddings of depth $K = 5$ or $K = 10$. In the right panel, we use $K = 5$, but use 4 quarters of data instead of a single quarter. The reported $R^2$ is an average of the quarterly cross-sectional $R^2$ values. The sample is from 2005.Q1 to 2020.Q4 and uses holdings data from 13F filings.



ongoing work, we are unbundling the 13F institutions to build the most comprehensive holdings data set feasible by combining 13F data and, when possible, the fund-level data.

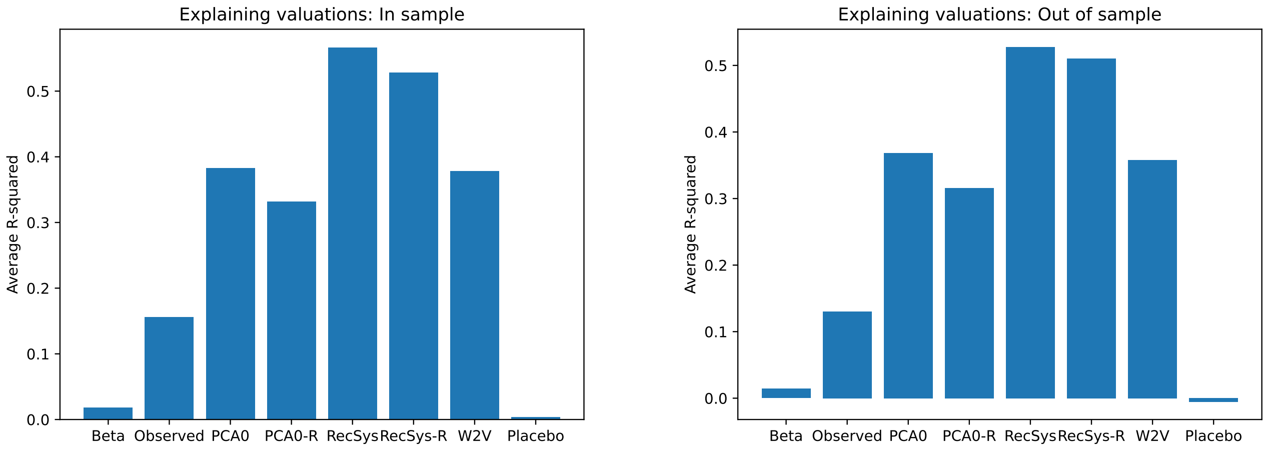
The main takeaway is that asset embeddings estimated from holdings data contain useful information to explain firms' valuations. While we can certainly expand the set of observable characteristics, and expand the models that we use to estimate asset embeddings, these first results illustrate that learned asset embeddings from holdings data can provide a valuable set of firm characteristics. In future work, we plan to also explore whether the embeddings that explain valuations beyond book equity primarily capture information about a firm's future cash flows or discount rates as in Vuolteenaho (2002) and Campbell and Vuolteenaho (2004).

### 5.1.3 Using asset embeddings to improve the practice of firm valuation

We think that asset embeddings could be useful for valuation, building on the analysis of Hommel et al. (2023). To value a firm (given its projected cash-flows), the basic textbook method is to use the cost of capital given by the CAPM, using one characteristics, the firm's beta. This is what MBA students learn, but the reality is that many of their teachers now are unconvinced that using the CAPM for the cost of capital is right – either in practice (as it gives poor results to explain existing valuations) or in theory (as the CAPM is a very special and antiquated model).[23] But what to do instead? Instead, people sometimes mention a multi-factor model (as in the Fama-

[23] See the Chicago Booth panel (April 18 2023), `https://www.kentclarkcenter.org/surveys/discount-rates/`

Figure 5: **Using asset embeddings to explain firm valuations: Data from funds** We compare the in-sample (left panel) and out-of-sample (right panel) performance of asset embeddings estimated using various methods. The bar “Beta” uses the CAPM beta, the bar “Observed” uses the characteristics in Koijen and Yogo (2019), “PCA0” uses PCA where the missing values are replaced by zeros, “PCA0-R” uses PCA, replaces the missing values by zeros, and replaces holdings by percentile ranks, “RecSys” uses the recommender system, “RecSys-R” uses the recommender system and replaces holdings by ranks, “W2V” uses the Word2Vec model, and “Placebo” are randomly-simulated embeddings. All models are estimated with the same dimensionality as the observed data ($K = 5$) and we use a single quarter of data. The reported $R^2$ is an average of the quarterly cross-sectional $R^2$ values. The sample is from 2006.Q1 to 2020.Q4 and uses holdings data from quarterly filings of mutual funds, exchange-traded funds, closed-end funds, and variable annuity funds.



French benchmark we used above), although they also tend to perform poorly for valuation purposes (Hommel et al., 2023). More frequently, they use "comparable" firms. But the choice of comparables is quite challenging. E.g., if Apple a tech company, a consumer good company etc? It has many rough comparables, and the comparability weight to give to each comparable firm is not clear a priori. Assets embeddings offer a systematic, data-driven way to find "comparables" — see the revealed tastes by investors for characteristics; and it also has a theoretical microfoundation (see Section 3).

To get the cost of capital of a firm, given its embeddings $x_a$, there are two methods. One, is given projected cashflows and its embedding-predicted valuation, we can infer a cost of capital. Second, we can see how embeddings predict returns (so that the risk premium at time $t$ is $\pi_{at} = \pi_x x_{at}$ for some vector $\pi_x$ that can be estimated. This yields a cost of capital.

Now, given a new firm, or a new project, on which we do not have holdings data yet, how to value its cost of capital? One approach can be the following. Given observable characteristics $x^o_{at}$, we can surmise its embedding $x_a$ via a regression, $\mathbb{E}\left[x_a | x^o_{at}\right] = b^x_{x^o} x^o_{at}$, where we estimate $b^x_{x^o}$. Then, we use also the estimate $\pi_{at} = \pi_x x_{at}$ (estimated from all the universe of firms) to get $\mathbb{E}\left[\pi_{at} | x^o_{at}\right] = \pi_x b^x_{x^o} x^o_{at}$. One can show that this is more precise than the rougher procedure of regressing directly excess returns on $x^o_{at}$,as this controls better for observables.[24] In sum, we think that asset embeddings could be used routinely to value firms and projects.

**Additional remarks on firm valuation with embeddings** One could use our predicted measure of valuation as a way to do return prediction: regressing return our measure of value $p^\delta_a := p^{\text{model}}_a - p_a$. More generally, the mean return could be predicted that $r_{a,t+1} = \beta x_{at} + \varepsilon_{a,t+1}$, or something more sophisticated using the investor embeddings, and their respective speed of adjustment.

When doing firm valuation, it is useful to have a measure of uncertainty about one's estimate. One way to do get model the log error variance as:

$$\ln var\left(p^\delta_a | x_a\right) = \beta^\sigma x_a \tag{9}$$

[24] If we run across firms $r_{a,t+1} - r_{ft} = \beta x^o_{at} + \varepsilon_{a,t+1}$, then estimated $\beta$ should be asymptotically equal to $\pi_x b^x_{x^o}$. The direct procedure is noisier than going through the embeddings—intuitively because the controlling for embeddings controls for more information, hence is more precise. Indeed, orthogonalize $x_a = \left(x^o_a, x^h_a\right)$ with $x^o_a$,$x^h_a$ uncorrelated. We have:

$$r_{a,t+1} - r_{ft} = \beta^o x^o_{at} + \beta^h x^h_{at} + e_{a,t+1}$$

So in the regression

$$r_{a,t+1} - r_{ft} = \beta^1 x^o_{at} + \varepsilon_{a,t+1}$$

we have $\varepsilon_{a,t+1} = \beta^h x^h_{at} + e_{a,t+1}$, which implies $\sigma_\varepsilon > \sigma_e$. Such a regression will, for large $T$, gives $\beta^1$ equal to the true value $\beta^o$. But to estimate $\beta$, the SE on the "direct procedure" to estimate $\beta^1$ is $SE^{\beta^1} = \frac{1}{\sqrt{T}} \frac{\sigma_\varepsilon}{\sigma_{x^o}}$, while the SE on procedure "via embeddings" is $SE^{\beta^o} = \frac{1}{\sqrt{T}} \frac{\sigma_e}{\sigma_{x^o}}$, which is smaller.

for a $\beta^{\sigma}$ to be estimated. One could also add extra derived measures, e.g. the "number of firms close to firm $a$ for dimension $k$", which might be operationalized as $n_a^k$ in:

$$n_a^k := \sum_b e^{-\frac{\left(x_a^k - x_b^k\right)^2}{\sigma^{k,2}}}, \qquad \sigma^{k,2} := var\left(x_c^k\right) \tag{10}$$

To predict the error in (9), one could append $n_a$ to $x_a$, i.e. replace $x_a$ by $\left(x_a', n_a'\right)'$.

## 5.2 The factor structure in returns

In the second application, we focus on the factor structure in returns, which has a long tradition in the asset pricing literature as it (i) is important for risk management and (ii) can yield factors that explain differences in expected returns (for behavioral or rational reasons). This is naturally a high hurdle as much effort has been devoted in the asset pricing literature to uncover sources of comovement and characteristics that forecast future returns.

### 5.2.1 Framework and implementation

We model the return in month $m$ of quarter $q$ as

$$r_{am(q)} = \delta_{m(q)} + x_{a,q-1}' f_{m(q)} + \epsilon_{am(q)},$$

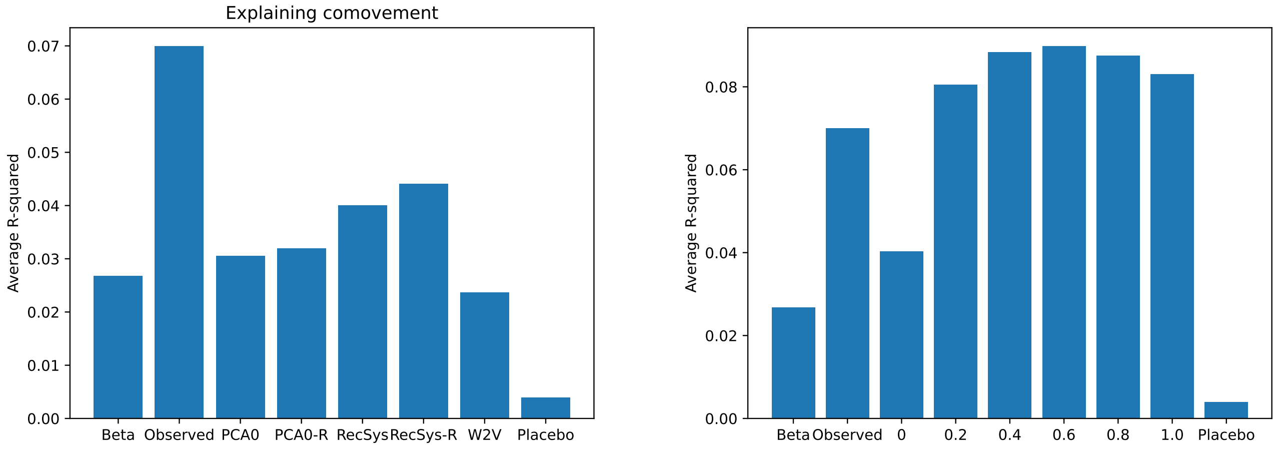
where $r_{am(q)}$ are the monthly returns and $f_{m(q)}$ are the factor realizations. We then record the monthly $R^2_{m(q)}$ and report the average across all months, $\frac{1}{M}\sum_m R^2_{m(q)}$.

We also implement supervised embeddings by using daily return data from quarter $q-1$ as the model data, see equation (1). We vary the importance of model data from $\kappa \in \{0, 0.2, .., 0.8, 1\}$. $\kappa = 0$ corresponds to a recommender system using holdings data alone, while $\kappa = 1$ corresponds to a recommender system using daily returns in the previous quarter. Both observable characteristics (the characteristics of the 5-factor Fama and French model in this case plus momentum) and a recommender system based on daily returns in the previous quarter are known to capture comovement in future returns. We compare those characteristics to asset embeddings that we learn from holdings data.

### 5.2.2 Empirical results

We report the main results in Figure 6. In the left panel, we compare the comovement explained by unsupervised asset embeddings, while in the right panel we report the results for the supervised recommender system. The left panel follows the same structure as Figure 2. In the right panel, we plot the $R^2$ for $\kappa \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$.

Figure 6: **Using asset embeddings to explain comovement** We compare comovement in returns using unsupervised asset embeddings in the left panel and supervised asset embeddings in the right panel using various methods. The bar "Beta" uses the CAPM beta, the bar "Observed" uses the characteristics in the 5-factor Fama and French model alongside momentum, "PCA0" uses PCA where the missing values are replaced by zeros, "PCA0-R" uses PCA, replaces the missing values by zeros, and replaces holdings by percentile ranks, "RecSys" uses the recommender system, "RecSys-R" uses the recommender system and replaces holdings by ranks, "W2V" uses the Word2Vec model, and "Placebo" are randomly-simulated embeddings. All models are estimated with the same dimensionality as the observed embeddings ($K = 6$) and we use a single quarter of data. The supervised model is estimated using daily returns in the same quarter as the holdings data. The reported $R^2$ is an average of the quarterly cross-sectional $R^2$ values. The sample is from 2005.Q1 to 2020.Q4 and uses holdings data from quartely filings of mutual funds, exchange-traded funds, closed-end funds, and variable annuity funds.



We find that the supervised asset embeddings are able to explain more of the cross-sectional variation in returns than the CAPM beta, and in particular the recommender system, but they fall short of matching the variation explained by the five characteristics of the 5-factor Fama and French model plus the momentum characteristic. This may not come as a surprise as much effort has been devoted to explaining comovement and, related, differences in stocks' average returns.

In the right panel, however, we see that supervised asset embeddings fare significantly better. $\kappa = 0$ is the same as the unsupervised recommender system based on log holdings, while $\kappa = 1$ corresponds to estimating a recommender system based on daily returns. We find, however, that the $R^2$ peaks at $\kappa = 0.6$, implying that there is additional information holdings that can be extracted as asset embeddings that yields the highest $R^2$ out of sample.

## 5.3 Asset similarity and substitution patterns

### 5.3.1 Framework

A growing literature explores models of asset demand. An important question is to determine which assets are close substitutes or not, as it affects the price impact of shift in demand.[25] However, which assets are close substitutes may depend on strategy of investors. For instance, Apple may be a close substitute to other technology companies for a technology fund, while is a close substitute to other large-cap stocks for a large-cap fund. The idea that asset embeddings depend on the context is what transformer models can capture, and we explore simple versions of such models in this section.

### 5.3.2 Implementation

To provide preliminary results, we report estimates for the AssetBERT model for a single quarter, 2019.Q4. We estimate a model with 4 layers, 2 attention heads, and an embedding depth of 8, 16, 32, 64, and 128.[26] This is a smaller model than the full model, and should therefore be viewed as a proof of concept, and future iterations of this paper will report larger-scale BERT models and compares their performance. Our first question is whether this simple implementation of the BERT model, trained on a masked language task, is able to identify meaningful variation in investors' portfolios.

### 5.3.3 Evaluating contextualized asset embeddings

The goal is to estimate the identity of a masked position in an investor's portfolio (or, depending on the application, which stock we expect an investor to trade). We can use information about all stocks in an investor's portfolio but not the masked stock.

In AssetBERT, we get a distribution over stocks for the masked position based on the other stocks in an investor's portfolio. We want something equivalent based on observed characteristics, $x_a^o$, but *not* using the characteristics of the masked stock. We propose the following algorithm (we focus on ranks but this logic also applies to holdings):

1. For each investor, fit ranks on embeddings, i.e. estimate $\lambda_{0i}, \lambda_{1i}$:

$$\rho_{ia} = \lambda_{0i} + \lambda_{1i}' x_a + \epsilon_{ia},$$

[25] We would enrich (5) into: $h_{ia} = c_i^h + \nu_{ia} + \left(1 - \zeta_i^{\perp}\right) p_a - \sum_{b \neq a} \zeta_{i,ab}^{;\perp} p_b$, where $\zeta_{i,ab}^{;\perp}$ is the cross-elasticity of demand between assets $a$ and $b$. When two assets are similar, the elasticity of substitution should be large (at least, under most microfoundations), so one could parametrize $\zeta_{i,ab}^{;\perp} = \zeta_{i,0}^{\perp} + \zeta_{i,1}^{\perp} d\left(x_a, x_b\right)^{-1}$, where $d\left(x_a, x_b\right)$ is a distance between embeddings, and $\zeta_0^{\perp}$, $\zeta_1^{\perp}$ are two parameters.

[26] The size of the feedforward network is four times the depth of the embedding, which is the ratio proposed in the original BERT model.

using data on all stocks except the masked position. We denote this set of unmasked stocks as $\mathcal{K}_i$.

2. We are interested in predicting a stock at rank $\rho$. We then define $\xi_{ia}(\rho) = |\rho - \lambda_{0i} - \lambda'_{1i}x_a|$ and construct a distribution over stocks as

$$\gamma_{ia} = \exp(\zeta\xi_{ia}(\rho))I(a \notin \mathcal{K}_i),$$

and

$$\mathbb{P}(\rho_{ia} = \rho \mid \mathcal{J}_{\rho i}) = \frac{\gamma_{ia}}{\sum_b \gamma_{ib}},$$

where $\zeta$ is a hyper parameter that is the same across investors, and $\mathcal{J}_{\rho i}$ is the information set used for rank $\rho$. We sum over all other stocks, other than those that are unmasked in the portfolio.

3. We can then compute, both for AssetBERT and observed characteristics the cross entropy of the masked words (in set $\mathcal{M}$)

$$-\frac{1}{N}\sum_{a\in\mathcal{M}} \log \mathbb{P}(\rho_{ia} = \rho \mid \mathcal{J}_{\rho i}).$$

Alternative, we can normalize the probabilities by the prediction of a randomly-selected position, to obtain the relative cross-entropy between the model, and a benchmark where stocks are picked randomly:

$$-\frac{1}{N}\sum_{a\in\mathcal{M}} \log \frac{\mathbb{P}(\rho_{ia} = \rho \mid \mathcal{J}_{\rho i})}{1/(A- \mid \mathcal{K}_i \mid)}, \tag{11}$$

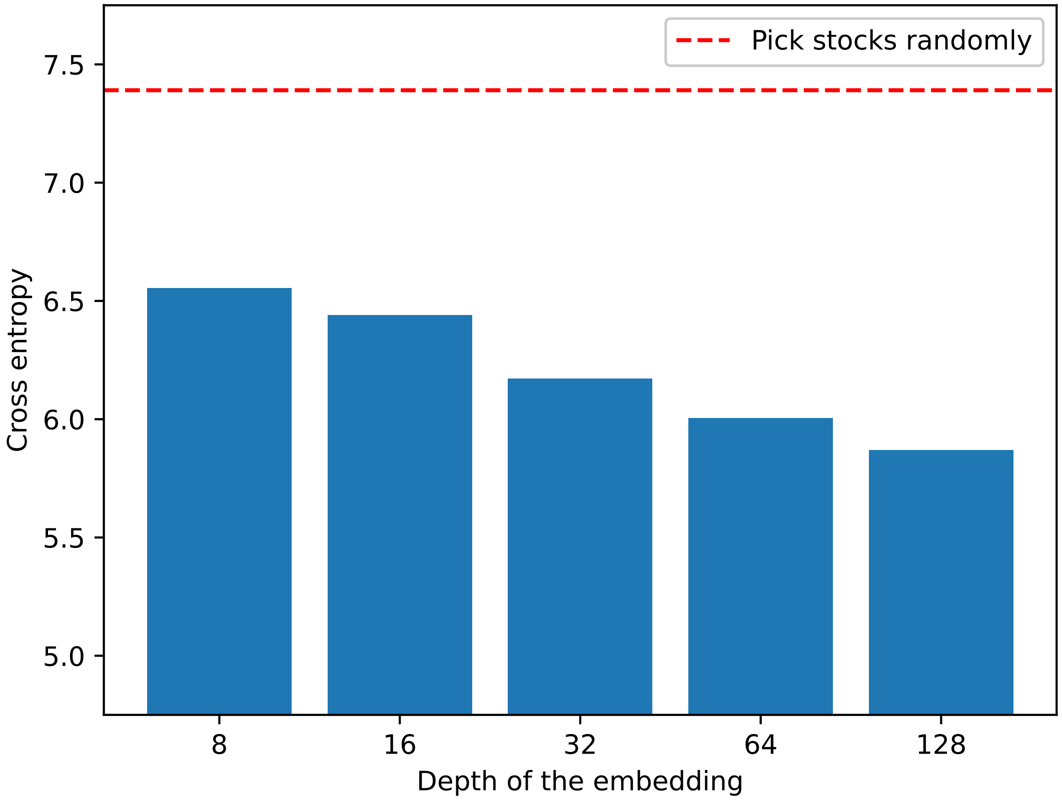
where $A$ denotes the total number of stocks.

#### 5.3.4 Empirical results

In Figure 7, we plot the cross entropy for different version of the AssetBERT model for embeddings depths of 8, 16, 32, 64, and 128. We also add the probability of stocks are selected randomly ($-\ln \frac{1}{1621}$ in our sample). The evaluation is done on a subset of the data that is not used in estimation.

We then run the evaluation as outlined in Section 5.3.3. We maximize $\zeta$ to give observable characteristics the best chance to predict the masked position. For the observed characteristics, we find that the relative cross entropy, see (11), equals -0.35. It implies that the likelihood ratio is given by $e^{0.35} = 1.4$. Using AssetBERT, the relative cross entropy is much lower at -1.67, which translates into a likelihood ratio of 5.3 compared to picking stocks randomly. Combining both results implies that AssetBERT performs 3.7 times better than observable characteristics to predict holdings an

Figure 7: Cross entropy for various specifications of the AssetBERT model



investor's portfolio. This suggests that transformer-based models are able to capture important substitution patterns among securities that are not well captured by observable characteristics.

# 6 Extensions

In this section, we discuss a series of extensions of the core ideas developed in this paper.

## 6.1 Unsupervised embeddings with context

One of the major breakthroughs in natural language processing (NLP) is to use transformer models to learn the context of words (Devlin et al. (2019)). In addition to the word embedding, the model also learns the position embedding and what the key words are to determine the meaning of a word in a sentence. This attention mechanism is central to many recent NLP breakthroughs.

## 6.2 Generalized supervised embeddings

In (21), we assume that the outcome variable is linear in the embedding vector, conditional on $d_{ta}$. This leads to a simple objective function in (22) that can be optimized easily by alternating least squares, as discussed in Section D.1.

However, supervised asset embeddings can also be used for non-linear objective functions. Suppose we are interested in minimizing the function $l(y_{ta}, x_a, d_{ta}; \theta^l)$ rather than $(y_{ta} - c_t - \beta_t' x_a)^2$. The optimization with respect to $\theta^l$, the parameters that are specific to the objective function, is done separately from the embeddings and replaces step 2 from the main algorithm (seen after (22)). The main challenge is updating the embeddings. To do so, we use a second-order expansion of function $\ell$ and proceed otherwise as in the main algorithm.[27]

To provide a concrete application, we can consider the parametric portfolio policies introduced by Brandt et al. (2009). They model the optimal portfolio as

$$w_a = w_a^{bm} + \theta' x_a,$$

where $w_a$ are the portfolio weights, $w_a^{bm}$ benchmark weights, and the term $\theta' x_a$ captures tilts away from the benchmark in the direction of characteristics. The parameter vector $\theta$ is then optimized

[27] Specifically, in step 3, we use in iteration $p$ a second-order approximation around $x_a = x_a^{(p-1)}$,

$$l_{\tilde{x}_a}(x_a^{(p-1)})' \Delta x_a^{(p)} + \frac{1}{2} \Delta x_a^{(p)\prime} l_{\tilde{x}_a \tilde{x}_a}(x_a^{(p-1)}) \Delta x_a^{(p)},$$

where we omit the constant and define $\Delta x_a^{(p)} = x_a^{(p)} - x_a^{(p-1)}$. This allows us to replace $B_a$ in step 3 and use the same algorithm otherwise. Of course, the function $l$ will need to satisfy some regularity conditions to ensure the convergence and uniqueness of the solution.

to maximize a CRRA utility objective function, which is the $\ell$ function above, times $-1$. We can then optimize over $\theta$ and estimate the optimal embeddings, $x_a$, to maximize an investor's utility.

## 6.3 Semi-parametric asset embeddings

We currently model asset embeddings as $x_a = (1, \tilde{x}_a')'$, where we estimate $\tilde{x}_a$. We can also allow for semi-parametric embeddings, where some of the components of the embedding vector depend on observable characteristics. The most obvious approach is to directly include observed characteristics, $x_a^o \in \mathbb{R}^O$,

$$x_a = (1, x_a^{o'}, \tilde{x}_a')'.$$

As we currently have a large number of characteristics, we can add a selection matrix

$$x_a^{o*} = G x_a^o,$$

where $G \in \mathbb{R}^{O^* \times O}$ and $O^* < O$. The matrix $G$ is to be estimated and dimension $O^*$ is a hyperparameter that can be determined alongside the depth of the embedding. The idea to reduce to optimally combine characteristics is not new to the literature, and also proposed by Kelly et al. (2019) in the context of factor models.

In sum, using semi-parametric asset embeddings, we strictly extend the information available in observable characteristics.

## 6.4 Alternative embedding data sets

Any data that has the dimensions in a period can be used as an embeddings data set. While we mostly focus on portfolio holdings and rebalancing (which across stocks and investors in any given quarter), we briefly discuss other data sets that can be used.

A key feature of embedding data is that it has two dimensions. In case of holdings data, the portfolio weights that vary across investors and across stocks. Other examples include daily returns or volume data within a quarter. The results in Section 3 provide conditions under which these data are informative as well. In future versions of this paper, we will explore combining different embedding data sets.

Beyond data on returns, volume, and portfolio holdings, it may be possible to use text data as well, for instance from earnings calls or 10-K filings across firms. We leave those extensions for future research.

## 6.5 Using investor embeddings

One can similarly recover investor embeddings — a vector $\lambda_{it}$ for each investor $i$ and time $t$. While there is detailed accounting data on firms, there is much less data on each investors. Hence, investor embeddings are likely to add a particularly high amount of information over readily-available data (e.g., institutional type, size, and activeness). In addition, it is clear that the characteristics of some assets do not change (e.g. the cash-flows of a given Treasury bond), so that brisk movements in their prices (controlling for e.g. the short term rate) must be attributed to changes in the investor embeddings (i.e., investors' demand or expectations), rather than the asset embeddings: changes in investor embeddings must be very important in the pricing process, at last for those assets. Having investor embeddings gives an opportunity to model investors, including with some behavioral features, such as speed and reactivity, or the clustering in different styles of investing. Hence, establishing and "explaining" investor embeddings seems like a productive way to make progress in structural finance.

## 6.6 Risk management and stress scenario

Central banks, regulators, and financial institutions could find asset and investor embeddings useful for risk management, including stress scenarios. Indeed, Generative AI generates new pictures, texts etc, from an initial prompts—e.g. via stable diffusion (Rombach et al. (2022)). Likewise, one could generate changes in asset and investors embeddings, that in turn lead to changes all market prices. This will generates useful stress scenarios, that include episodes never seen in real life, but still plausible generalizations from past scenarios.

## 6.7 Using text data in addition to portfolio data

One can also use text data, e.g. from the newspapers, or from the firms' filings or earnings calls (see Hassan et al. (2019), Bybee et al. (2023), and Chen and Sarkar (2020)). Those extra $x_a^{\text{text}}$ can be merged with our $x_a$ from holdings, this way augmenting the quality of the embeddings (by how much is an interesting research question). One could also use text data to see better the "meaning" of a characteristics $k$, e.g. by selecting the stocks most representative of it, and how see what words they are particularly associated with at a given date.

## 6.8 Generative portfolios: Forming factor mimicking portfolios without return data

Another application is "generative portfolios". For instance, suppose that we want to have a "Covid" factor, doing badly when Covid is strong, and form factor mimicking portfolio. One starting point is to be long Cruise Carnival and short Zoom, as they are exemplar stock for sensitivity to Covid. But this portfolio formed of only two stocks has lots of idiosyncratic risk. So, one uses asset embeddings to find stocks "similar" to Zoom and Cruise Carnival, and add them to the portfolio, and diversifying the idiosyncratic risk. Formally, we might search for assets $a$ with a high value of $V_a := s(x_a, x_{\mathrm{Cruise}}) - s(x_a, x_{\mathrm{Zoom}})$, where $s(x_a, x_b) = \frac{x_a \cdot x_b}{\|x_a\| \|x_b\|}$ is the cosine similarity.[28] We put the assets with a high positive value in the long leg of the portfolio, assets with a high negative value on the short leg. This way, we get a much better factor mimicking portfolio.[29]

One could even roll the procedure one step back. If we link the model with text, we can ask for an "inflation factor." It would scan text data (e.g.Wall Street Journal) and find some stocks mentioned in those articles. We then use AssetBERT to complement the portfolio. Then we can go from an idea in words to a factor mimicking portfolio.

## 6.9 Further research questions using embeddings

**Benchmarked competition for economics** One impetus for progress in machine learning was the use of "benchmark tests" – periodic competitions in which different teams try perform tasks, e.g. classify images, or predict the 3-dimensional structure of a protein. The most famous are the Kaggle competitions[30] and we are proposing here a form of "economics Kaggle competition".

In this paper, we propose tasks that would be ripe for such challenge, a procedure to meet it: (i) Predicting valuations of a number of "masked" firms, given (as we have here), the valuation of other firms on the same date, and investor holdings. (ii) Predicting "masked" holdings—given a partial data on holdings data, one has to predict the holdings of a given set of investors.

One could imagine a competition in which researchers send in advance their code one quarter in advance, and the success is evaluated every quarter when the new data (on valuations and holdings).[31]

Another such challenge would include, based on a research team's favorite way to build embeddings: (iii) given few macro variables (e.g. interest rates) in addition to embeddings, predict a firms' hiring and investment; (iv) predict the matrix of ex post correlation between assets.

[28] Ideally, this would be a weighted similarity, weighing some dimensions more than others.

[29] Obviously there are many variants on that theme: for instance, one might add more exemplar stocks (e.g. add American Airlines to Cruise Carnival positive exemplar), and add bonus points for e.g. liquid assets.

[30] `www.kaggle.com/competitions`. There is also the General Language Understanding Evaluation (GLUE) competition.

[31] In a sense, economics already has some pre-existing benchmark hurdles for calibrated theories: for instance, matching various moments in the business cycle (as in Kydland and Prescott (1982)), various moments of returns in a macro-finance model (as in Campbell and Cochrane (1999)), or having small errors in the alpha of various test portfolios (as in Fama and French (1992)). But returns are so noisy, and big macro variables so few, that a new quarter of data is not very diagnostic to update the merit of models. Whereas, when the task is to predict partially masked valuations and holdings, as those are less noisy, the tests can be much more revealing, and the competition interesting.

**Understanding non-financial actions by firms** Those embeddings should be able to predict also real (i.e. non-financial) actions by firms, e.g. hiring and investment. It would be interesting to see how large is that explanatory power, and which might lead to a better understanding of investment—for instance, in some periods, investors reward investment and growth, in some others, they reward profits and dividends. Asset and investor embeddings ought to be useful to measure that, which then might inform theories about the dynamics.

**Understanding intangible information** Suppose that we decompose the embedding vector (after linear transformation) as $x_a = \left(x_a^o, x_a^h\right)$ where $x_a^o$ is the vector of observable characteristics (e.g. from accounting data, past growth, etc.) and $x_a^h$ the vector of hidden or intangible characteristics. For instance $x_a^h$ might include "number of visit to your website" during the internet stock boom of the late 1990s, or "use of GPUs" during the AI boom of 2023, and at all times "sensitivity to the dollar". It would be interesting to see how investors react to observable vs hidden characteristics.

**IPOs** Embeddings should also be useful to understand IPO returns. Suppose that one knew the composition of investors ("the book") right before the IPO. Then, in principle, one could gain extra prediction on the IPO returns.

# 7 Conclusion

We introduce the concept of asset embeddings and argue, theoretically and empirically, that portfolio holdings are a natural source of embedding data. Just as documents are useful to uncover word structures, investors' holdings reveal asset structures or embeddings. Embeddings can be used to complement the standard set of firm characteristics that are ubiquitously used in finance and economics.

Our current applications are in equity markets, but the same ideas can be applied to other asset classes such as fixed income markets, commodities, and currency markets as holdings data are becoming widely available across asset classes and geographies, including for US households (Gabaix et al. (2022)).

Our conceptual insight that holdings data contain key information about asset characteristics or embeddings also provides a natural connection between finance (and economics more broadly) and recent methodological advances in the fields of machine learning and artificial intelligence that we are exploring in ongoing work.

# References

**Balasubramaniam, Vimal, John Y. Campbell, Tarun Ramadorai, and Benjamin Ranish**, "Who Owns What? A Factor Model for Direct Stockholding," *Journal of Finance*, 2023, *78* (3), 1545–1591.

**Betermier, Sebastien, Laurent E. Calvet, Samuli Knüpfer, and Jens Soerlie Kvaerner**, "What Do the Portfolios of Individual Investors Reveal About the Cross-Section of Equity Returns?," 2022. Working paper.

**Brandt, Michael W., Pedro Santa-Clara, and Rossen Valkanov**, "Parametric Portfolio Policies: Exploiting Characteristics in the Cross-Section of Equity Returns," *Review of Financial Studies*, 2009, *22* (9), 3411–3447.

**Bryzgalova, Svetlana, Sven Lerner, Martin Lettau, and Markus Pelger**, "Missing Financial Data," 2022. Working paper.

**Bryzgalova, Svetlana, Victor DeMiguel, Sicong Li, and Markus Pelger**, "Asset-Pricing Factors with Economic Targets," *Working paper*, 2023.

**Bybee, Leland, Bryan Kelly, and Yinan Su**, "Narrative Asset Pricing: Interpretable Systematic Risk Factors from News Text," *The Review of Financial Studies*, 2023.

**Campbell, John Y. and John H. Cochrane**, "By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior," *Journal of Political Economy*, 1999, *107* (2), 205–251.

**Campbell, John Y. and Tuomo Vuolteenaho**, "Bad Beta, Good Beta," *American Economic Review*, 2004, *94* (5), 1249–1275.

**Chen, Jiafeng and Suproteem K. Sarkar**, "A Semantic Approach to Financial Fundamentals," 2020. Working paper.

**Chen, Luyang, Markus Pelger, and Jason Zhu**, "Deep learning in asset pricing," *Management Science*, 2023.

**Creal, Drew, Siem-Jan Koopman, and Andre Lucas**, "Generalized Autoregressive Score Models with Applications," *Journal of Applied Econometrics*, 2013, *28* (5), 777–795.

**Daniel, Kent, Mark Grinblatt, Sheridan Titman, and Russ Wermers**, "Measuring Mutual Fund Performance with Characteristic-Based Benchmarks," *Journal of Finance*, 1997, *52* (3), 1035–1058.

**Deerwater, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman**, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, 1990, *41* (6), 391–407.

**DeMiguel, Victor, Alberto Martin-Utrera, Francisco J. Nogales, and Raman Uppal**, "A Transaction-Cost Perspective on the Multitude of Firm Characteristics," *Review of Financial Studies*, 2020, *33*, 2180–2222.

**Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova**, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of NAACL-HLT 2019*, 2019.

**Dolphin, Rian, Barry Smyth, and Ruihai Dong**, "Stock Embeddings: Learning Distributed Representations for Financial Assets," *Working paper*, 2022.

**Dubinsky, Andrew, Michael Johannes, Andreas Kaeck, and Norman J Seeger**, "Option Pricing of Earnings Announcement Risks," *Review of Financial Studies*, 2019, *32* (2), 646–687.

**Ducharme, Yoshua Bengio Rejean, Pascal Vincent, and Christian Jauvin**, "A Neural Probabilistic Language Model," *Journal of Machine Learning Research*, 2003, *3*, 1137–1155.

**Dumais, Susan T, George W Furnas, Thomas K Landauer, Scott Deerwester, and Richard Harshman**, "Using latent semantic analysis to improve access to textual information," in "Proceedings of the SIGCHI conference on Human factors in computing systems" 1988, pp. 281–285.

**Fama, Eugene F. and Kenneth R. French**, "The Cross-Section of Expected Stock Returns," *Journal of Finance*, 1992, *47* (2), 427–465.

**Feng, Guanhao, Stefano Giglio, and Dacheng Xiu**, "Taming the Factor Zoo: A Test of New Factors," *Journal of Finance*, 2020, *75*, 1327–1370.

**Freyberger, Joachim, Andreas Neuhierl, and Michael Weber**, "Dissecting Characteristics Nonparametrically," *Review of Financial Studies*, 2020, *33*, 2326–2377.

**Freyberger, Joachim, Bjorn Hoppner, Andreas Neuhierl, and Michael Weber**, "Missing Data in Asset Pricing Panels," 2022. Working paper.

**Gabaix, Xavier and Ralph SJ Koijen**, "Granular instrumental variables," Working Paper 28204, National Bureau of Economic Research December 2020.

**Gabaix, Xavier and Ralph SJ Koijen**, "In search of the origins of financial fluctuations: The inelastic markets hypothesis," *NBER Working Paper No. 28967*, 2022.

**Gabaix, Xavier, Ralph S.J. Koijen, Federico Mainardi, Sangmin Oh, and Motohiro Yogo**, "Asset Demand of U.S. Households," *Working paper*, 2022.

**Gu, Shihao, Bryan Kelly, and Dacheng Xiu**, "Empirical asset pricing via machine learning," *The Review of Financial Studies*, 2020, *33* (5), 2223–2273.

**Gunasekar, Suriya, Yi Zhang, Jyoti Aneja, Caio Cesar Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sebastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li**, "Textbooks Are All You Need," *Working paper, Microsoft Research*, 2023.

**Hansen, Lars Peter and Scott F Richard**, "The role of conditioning information in deducing testable restrictions implied by dynamic asset pricing models," *Econometrica: Journal of the Econometric Society*, 1987, pp. 587–613.

**Hassan, Tarek A, Stephan Hollander, Laurence Van Lent, and Ahmed Tahoun**, "Firm-level political risk: Measurement and effects," *The Quarterly Journal of Economics*, 2019, *134* (4), 2135–2202.

**Hommel, Nicolas, Augustin Landier, and David Thesmar**, "Corporate Valuation: An Empirical Comparison of Discounting Methods," 2023. Working paper.

**Jurafsky, Dan and James H. Martin**, *Speech and Language Processing (3rd ed. draft)* 2023.

**Kelly, Bryan and Dacheng Xiu**, "Financial Machine Learning," 2023. Working paper.

**Kelly, Bryan T., Seth Pruitt, and Yinan Su**, "Characteristics are covariances: A unified model of risk and return," *Journal of Financial Economics*, 2019, *134*, 501–524.

**Koijen, Ralph SJ and Motohiro Yogo**, "A demand system approach to asset pricing," *Journal of Political Economy*, 2019, *127* (4), 1475–1515.

**Koijen, Ralph SJ, Robert J Richmond, and Motohiro Yogo**, "Which Investors Matter for Equity Valuations and Expected Returns?," *Working Paper*, 2022.

**Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh**, "Interpreting Factor Models," *Journal of Finance*, 2018, *73* (3), 1183–1223.

**Kozak, Serhiy, Stefan Nagel, and Shrihari Santosh**, "Shrinking the cross-section," *Journal of Financial Economics*, 2020, *135*, 271–292.

**Kydland, Finn E. and Edward C. Prescott**, "Time to Build and Aggregate Fluctuations," *Econometrica*, 1982, *50* (6), 1345–1370.

**Lettau, Martin and Markus Pelger**, "Factors That Fit the Time Series and Cross-Section of Stock Returns," *Review of Financial Studies*, 2020, *33*, 2274–2325.

**Madhavan, Ananth, Aleksander Sobczyk, and Andrew Ang**, "What Happens With More Funds Than Stocks?," *Journal of Investment Management*, 2021, *19* (2), 4–28.

**Michel, Paul, Omer Levy, and Graham Neubig**, "Are Sixteen Heads Really Better than One?," *33rd Conference on Neural Information Processing Systems*, 2019.

**Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean**, "Distributed Representations of Words and Phrases and their Compositionality," *NIPS*, 2013, pp. 3111–3119.

**Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean**, "Efficient Estimation of Word Representations in Vector Space," *ICLR Workshop Papers*, 2013.

**Nagel, Stefan**, *Machine learning in asset pricing*, Vol. 8, Princeton University Press, 2021.

**Pennington, Jeffrey, Richard Socher, and Christopher D. Manning**, "GloVe: Global Vectors for Word Representation," *Conference on Empirical Methods on Natural Language Processing*, 2014.

**Prince, Simon J.D.**, *Understanding Deep Learning*, MIT Press, 2023.

**Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever**, "Improving Language Understanding by Generative Pre-Training," *OpenAI*, 2018.

**Rombach, Robin, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer**, "High-resolution image synthesis with latent diffusion models," in "Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition" 2022, pp. 10684–10695.

**Ross, Stephen A.**, "The Arbitrage Theory of Capital Asset Pricing," *Journal of Economic Theory*, 1976, *13* (3), 341–360.

**Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin**, "Attention Is All You Need," *31st Conference on Neural Information Processing Systems*, 2017.

**Vuolteenaho, Tuomo**, "What Drives Firm-Level Stock Returns?," *Journal of Finance*, 2002, *57* (1), 233–264.

# A Appendix: Portfolio holdings as embedding data

## A.1 Equilibrium asset prices

Starting from the model in (5) and (6), we solve for asset prices by imposing market clearing. As we normalized the number of shares of asset $a$ to 1:

$$\sum_i H_{ia} = P_a.$$

To obtain closed-form solutions, we linearize the market clearing condition around a long-run equilibrium of demand and valuations, $(\bar{h}_{ia}, \bar{p}_a)$. We denote the long-run portfolio share by $\bar{w}_a$, $\sum_a \bar{w}_a = 1$, and log dollar value of assets by $\bar{a}_i$, implying $\bar{h}_{ia} = \bar{a}_i + \ln \bar{w}_a$. This implies for long-run valuations $\sum_i \bar{H}_{ia} = \bar{P}_a$, or, equivalently, $\bar{p}_a = \ln \bar{w}_a + \ln \sum_i \exp(\bar{a}_i)$. If we linearize the first-order condition and solve for prices, we obtain[32]

$$p_a = c^p + \frac{\nu_{S^a a}}{\zeta_{S^a}} = c^p + \frac{1}{\zeta_{S^a}} \lambda^{\nu\prime}_{S^a} x_a + u_{S^a a}, \tag{12}$$

where $X_S := \sum_i S_i X_i$ is the average of a variable $X_i$, weighing by importance $S_i$ of investor $i$; $S_i^a = \frac{\exp(\bar{h}_{ia})}{\sum_j \exp(\bar{h}_{ja})}$ is a specific measure that of importance, namely the fraction of asset $a$ held by investor $i$, and $c^p$ is an unimportant constant. In what follow, to simplify the notations, and replace $S_i^a$ by $S_i$.

This shows that we can recover embeddings from valuations and returns. However, to do so, we need a longer time period and assume that the loadings, $\frac{1}{\zeta_S}\lambda^{\nu}_S$, and the embeddings, $x_a$, do not change over time. Holdings data do not require this restriction, as we will show next.

## A.2 Equilibrium portfolio holdings

We substitute (12) into (5) and obtain

$$h_{ia} = \phi_i^h + \phi_a^h + \lambda_i' x_a + \epsilon_{ia}, \tag{13}$$

[32] We start from a linear approximation of the market clearing equation

$$\sum_i \exp(\bar{h}_{ia}) + \sum_i \exp(\bar{h}_{ia})(h_{ia} - \bar{h}_{ia}) = \exp(\bar{p}_a) + \exp(\bar{p}_a)(p_a - \bar{p}_a),$$

implying $p_a = \bar{p}_a + h_{Sa} - \bar{h}_{Sa} = \ln \sum_i \exp(\bar{a}_i) - \bar{a}_S + c_S^h + (1 - \zeta_S)p_a + \nu_{Sa}$ and thus

$$p_a = \frac{\ln \sum_i \exp(\bar{a}_i) - \bar{a}_S + c_S^h}{\zeta_S} + \frac{\nu_{Sa}}{\zeta_S}.$$

where $\phi_i^h = (1-\zeta_i)\, c^p$, $\phi_a^h = \frac{1}{\zeta_S}\left(u_{Sa} + \lambda_S^{\nu\prime} x_a\right)$, $\epsilon_{ia} = u_{ia} - \frac{\zeta_i}{\zeta_S} u_{Sa}$, and $\lambda_i = \lambda_i^\nu - \frac{\zeta_i}{\zeta_S}\lambda_S^\nu$.

Hence, the recoverable investor embedding from holdings data is not $\lambda_i^\nu$ but its close cousin $\lambda_i = \lambda_i^\nu - \frac{\zeta_i}{\zeta_S}\lambda_S^\nu$: it captures the differential sensitivity of investor $i$ to the asset characteristics $x_a$ compared to the average sensitivity of the other investors, which is the $\lambda_S^\nu$ term. This includes the differential price elasticity, which is captured by $\frac{\zeta_i}{\zeta_S}$. If all investors were identical (up to size) we would have $\lambda_i = 0$.

When demand elasticities are the same across investors, $\zeta_i = \zeta$, then then term $\frac{\zeta_i}{\zeta_S} u_{Sa}$ in $\epsilon_{ia}$ only varies across assets and is part of $\phi_a^h$ instead.[33] We then obtain a pure factor model. Otherwise, there is some small contamination from investors trading against idiosyncratic demand shocks of other investors.

The main insight is that as long as investors disagree on how asset embeddings affect risk and expected return, or use asset embeddings in building portfolios for non-pecuniary reasons, portfolio holdings across investors are perfectly suited to extract embeddings. Investors who take bold bets and disagree more are particularly well suited to extract asset embeddings.

## A.3 Returns, volume, and portfolio rebalancing

In addition to portfolio holdings, rebalancing is informative as well. This is particularly true if some holdings are slow to adjust and rebalancing provides more accurate information. We first study quarterly returns, $r_{aq} = \Delta p_{aq}$ — we omit dividends for simplicity:

$$r_{aq} = \frac{1}{\zeta_S}\Delta\left(\lambda_{Sq}^{\nu\prime} x_{aq}\right) + \Delta u_{Saq}.$$

If embeddings are relatively stable over time, $\Delta x_{aq} \simeq 0$, then $\Delta\left(\lambda_{Sq}^{\nu\prime} x_{aq}\right) = \left(\Delta\lambda_{Sq}^{\nu}\right)' x_{a,q-1}$ and

$$r_{aq} = \frac{1}{\zeta_S}\left(\Delta\lambda_{Sq}^{\nu}\right)' x_{a,q-1} + \Delta u_{Saq}. \tag{14}$$

This implies that factor models of returns can also be used to uncover embeddings that are fairly stable over time. Instead of using quarterly returns, we can use daily returns (or, indeed, volume data) as embedding data, having provided a micro foundation here.

Following the same logic, we can derive an expression for investors' portfolio rebalancing

$$\Delta h_{iaq} = \Delta\phi_{iq}^h + \Delta\phi_{aq}^h + \Delta\lambda_{iq}' x_{a,q-1} + \Delta\epsilon_{iaq}. \tag{15}$$

As is clear from (14) and (15), we can also estimate asset embeddings based on portfolio rebalancing, $\Delta q_{ia} = \Delta h_{ia} - r_{aq}$.

[33] Also, the term $-\frac{\lambda_S^\nu}{\zeta_S}\zeta_i = -\lambda_S^\nu$ in $\lambda_i$ no longer varies across investors or assets, and the term $-\lambda_S^{\nu\prime} x_a$ therefore also folds into $\phi_a^h$.